DISS ETH NO. 28067

Investigation of the protein correlated motion and allostery based on NMR structural ensembles

A thesis submitted to attain the degree of

DOCTOR OF SCIENCE OF ETH ZURICH
(Dr. Sc. ETH Zurich)

presented by

Dzmitry Ashkinadze

M.Sc. ETH Zürich

born on *15.05.1994*

citizen of Belarus

accepted on the recommendation of:

Prof. Dr. Roland Riek	*(Examiner)*
Prof. Dr. Gunnar Jeschke	*(Co-examiner)*
Prof. Dr. Peter Güntert	*(Co-examiner)*

2021

«…

Do what you should, and come what may

…»

Marcus Aurelius, 121AD

Acknowledgments

I would like to thank my direct supervisor, Prof. Roland Riek, for guiding me throughout my Ph.D., for continuous support in my various projects and for his intuition that helped me find significance. I am grateful for the opportunity to travel to Japan and participate in a collaborative nanodiamond project with Prof. Shirakawa and Dr. Segawa. I am also grateful that I was given a second chance and a computational project for a personal reason. Despite his principal investigator status, he always had time to teach me about protein assignment and shape together my developing projects. His approach in the validation of multi-state protein structures acted as a blueprint for most of the contributions that I was able to create during my Ph.D.

I would like to thank Prof. Peter Güntert for his involvement and continuous support in all of my structural biology projects. His vast expertise with protein structure calculation and program CYANA of his creation helped me to improve my understanding of both what is a good quality protein structure and what is valued by researchers in the whole structural biology field.

I would like to thank Prof. Gunnar Jeschke for accepting to be my co-examiner and Prof. Beat H. Meier to be a chairperson during my Ph.D. examination.

I would like to thank Dr. Harindranath Kadavath for his expertise in the field of NMR and protein allostery. On numerous occasions, he assisted me with NMR measurements or even performed them for our shared projects. Our discussions allowed me to improve my scientific writing as well as my biological understanding of protein allostery and NMR.

I would like to thank Dr. Piotr Klukowski for his machine learning expertise that allowed me to deepen my understanding of computational algorithms and possibly find my future job.

I would like to thank Prof. Shirakawa for the hospitality and hosting of my research in Japan.

I would like to thank Dr. Lukas Frey who introduced me to the protein purification techniques and supervised my laboratory lipid nanodisc project during the first year of

my Ph.D. and Dr. Jason Greenwald who knew how to fix problems that I and Lukas could not solve.

I would like to thank the rest of the Riek group for the discussions and good working atmosphere.

I would like to thank my family for their moral support and words of encouragement. The financial support of my parents, their love, and their persistent belief in my abilities enabled me to fulfill my dream of going to study abroad and becoming a researcher. My wife, Anastasia who is a doctoral student herself shares with me all of the personal and professional ups and downs, together we are more than a sum of us and it would not be an understatement that all of my Ph.D. achievements are her achievements as well. My late grandfather Vladimir Reutsky was a pronounced academic researcher and a professor in the field of tree biology, he gave me the idea to go and study at the ETHZ. He was always an example of a true researcher to me and he continued to amaze me with the sheer range of his scientific curiosity and imagination. Once he sent a proposal to reverse climate change to a billionaire Richard Branson. He wanted to convert the Sahara Desert green with help of the hydroponic system developed by him and his colleagues for the submarines. As an alchemist, he also treated rare skin diseases with help of nuts, alcohol, and ginseng root. My Ph.D. thesis is dedicated to him. He would be proud.

Translation to Russian of the last paragraph:

“Отдельное спасибо моей семье за их постоянную помощь и поддержку. Благодаря финансовой помощи моих родителей, их любови и непоколебимой вере в мои возможности мне удалось поехать заграницу и стать настоящим ученым. Вместе с моей женой, Анастасией, которая заканчивает аспирантуру, как и я, мы делим все личные и профессиональные свершения и неудачи, вместе мы сильнее чем по отдельности, и я с уверенностью могу сказать, что все мои достижения, которые я смог добиться за свою аспирантуру были бы невозможны без нее. Мой дедушка, Владимир Григорьевич Реуцкий был выдающимся ученым, академиком и профессором биологии деревьев Белорусской Академии наук, именно он подал мне идею про обучение в Швейцарском Техническом Институте Цюриха. Он всегда был для меня примером настоящего ученого. Я не раз удивлялся его научному

любопытству и масштабу его воображения включая тот случай, когда он послал свое предложение по борьбе с глобальным потеплением миллиардеру Ричарду Бренсену. Он предлагал озеленить Сахару с помощью гидропоники, созданной им и его коллегами для подводных лодок. Также он как древний алхимик успешно лечил редкие кожные заболевания с помощью орехов, водки и корня женьшеня. Ему я посвящаю свою аспирантскую работу, он бы гордился.”

Abstract

Protein dynamics and protein correlated motion are the key for understanding of most mechanisms behind target recognition, enzymatic activity and signal transduction. Recent advances based on exact Nuclear Overhauser Effect (eNOE) developed by the group of Prof. Riek in the field of protein structure determination using liquid-state nuclear magnetic resonance (NMR) enable the elucidation of multi-state protein conformations of atomic resolution that sample protein conformational space. However, so far eNOE approach was applied to the limited number of proteins mostly by the group of Prof. Riek.

As an extension of eNOE dataset the protein allostery, correlated motion and ligand binding mode of the protein PDZ2 was investigated and two-state protein structures were calculated for both free form and bound to the RA-GEF2 peptide with eNOEs. Apo-holo structural rearrangements allowed to reconstruct protein allostery that validates previously published allosteric interactions from groups of Ranganathan and Lee [1, 2]. A novel allosteric conformational preselection step was detected and apo protein states were identified to be “open” and “closed” due to the obstruction of the binding site by sidechains of residues Lys38 and Lys72.

In order to quantify the correlated motion involved in the conformational selection allosteric mechanism of the PDZ2 multi-state structure an automated and unbiased method PDBcor was developed. PDBcor is a software for the detection and analysis of correlated motions from experimental multi-state protein structures using torsion angle and distance statistics that does not require any structure superposition. Clustering of protein conformers allows us to extract correlations with high sensitivity in the form of mutual information based on information theory. With PDBcor we elucidated correlated motion in the PDZ2 domain, WW domain of PIN1, the protein GB3, and the enzyme cyclophilin in line with reported findings. As a guide for the interpretation of PDBcor results, we provide a series of protein structure ensembles that exhibit different levels of correlation, including non-correlated, locally correlated, and globally correlated ensembles.

Correlations extracted with PDBcor can be utilized in subsequent assays including NMR multi-state structure optimization and validation. So far, NMR derived

multi-state structures were typically evaluated by means of visual inspection of structure superpositions, target function values that quantify the violation of experimented restraints and root-mean-square deviations (RMSD) that quantify similarity between conformers. As an alternative or complementary approach, we present here the use of a recently introduced structural correlation measure, PDBcor, that quantifies the clustering of protein states as an additional measure for multi-state protein structure analysis. It can be used for various assays including the validation of experimental distance restraints, optimization of the number of protein states, identification of key distance restraints, NOE network analysis and semiquantitative analysis of the protein correlation network. We present applications for the final quality analysis stages of typical multi-state protein structure calculations.

Extensive testing of the new tools developed for extraction and investigation of protein correlated motion in form of structural correlations led us to the discovery that even conventional single-state liquid NMR protein structures that by design average out all state-dependent information contain valid structural correlations. Here we provide a potential mechanism for retention of such structural correlations on example of minimal systems and validate it with synthetically prepared data. We also show valid structural correlations on an example of experimental single-state liquid NMR structure of the protein cyclophilin A. Furthermore, we present structural correlation results for the whole PDB database and evidence that suggests that structural correlations of the single-state liquid NMR protein structures overlap with protein allosteric sites and might give insights into protein allostery.

Zusammenfassung

Proteindynamik und Proteinbewegung sind essenziel zum Verständnis der meisten Mechanismen hinter der Zielerkennung, der enzymatischen Aktivität und der Signalübertragung. Der exakte Nuclear Overhauser Effekt (eNOE), der in der Gruppe von Prof. Riek für die Proteinstrukturbestimmung mittels Flüssigzustands-NMR entwickelt wurde, ermöglicht die Aufklärung von Mehrzustands-Proteinkonformationen mit atomarer Auflösung, die den Konformationsraum von Proteinen abtasten. Bisher wurde der eNOE-Ansatz nur auf eine begrenzte Anzahl von Proteinen angewendet.

Als Erweiterung des eNOE-Datensatzes wurden die Proteinallosterie, die korrelierte Bewegung und der Ligandenbindungsmodus des Proteins PDZ2 untersucht und Zweizustands-Proteinstrukturen sowohl für die freie Form als auch für die Bindung an das RA-GEF2-Peptid mit eNOEs berechnet. Apo-Holo-Strukturumlagerungen ermöglichten die Rekonstruktion der Proteinallosterie, die zuvor veröffentlichte allosterische Wechselwirkungen von Gruppen von Ranganathan und Lee validiert [1, 2]. Ein neuer allosterischer Konformationsvorselektionsschritt wurde entdeckt und Apo-Proteinzustände wurden als „offen“ und „geschlossen“ aufgrund der Blockierung der Bindungsstelle durch die Seitenketten der Reste Lys38 und Lys72 identifiziert. Proteinkorrelierte Bewegung wurde mit Temperaturtitrationsexperimenten untersucht. Allosterischen Reste überlappten weitgehend mit Resten, die an der Proteinkorrelierten Bewegung beteiligt sind.

Um die korrelierte Bewegung aus der PDZ2-Mehrzustandsstruktur mit atomarer Auflösung zu quantifizieren, wurde eine automatisierte und unverzerrte Methode PDBcor entwickelt. PDBcor ist eine Software zur Erkennung und Analyse korrelierter Bewegungen aus experimentellen Mehrzustands-Proteinstrukturen unter Verwendung von Torsionswinkel- und Distanzstatistiken, die keine Strukturüberlagerung erfordert. Das Klustern von Proteinkonformeren ermöglicht es uns, Korrelationen in Form von gegenseitiger Information basierend auf der Informationstheorie zu extrahieren. Mit PDBcor haben wir die korrelierte Bewegung in der PDZ2-Domäne, der WW-Domäne von PIN1, dem Protein GB3 und dem Enzym Cyclophilin in Übereinstimmung mit den berichteten Ergebnissen aufgeklärt. Als Leitfaden für die Interpretation der PDBcor-Ergebnisse zeigen wir eine Reihe von Proteinstruktur-Ensembles die unterschiedliche

Korrelationsniveaus aufweisen, einschließlich nicht-korrelierter, lokal korrelierter und global korrelierter Ensembles.

Mit PDBcor extrahierte Korrelationen können auch in der NMR-Mehrzustandsstrukturoptimierung und -validierung. Bisher wurden NMR-abgeleitete Mehrzustandsstrukturen typischerweise durch visuelle Inspektion von Strukturüberlagerungen, Targetfunktionswerten, die die Verletzung experimenteller Beschränkungen quantifizieren, und quadratischen Mittelwertabweichungen (RMSD), die die Ähnlichkeit zwischen Konformeren quantifizieren, bewertet. Als alternativen oder ergänzenden Ansatz präsentieren wir hier die Verwendung eines kürzlich eingeführten strukturellen Korrelationsmaßes, PDBcor, das die Clusterbildung von Proteinzuständen als zusätzliches Maß für die Mehrzustands-Proteinstrukturanalyse quantifiziert. Es kann für verschiedene Assay verwendet werden, einschließlich der Validierung experimenteller Distanzbeschränkungen, der Optimierung der Anzahl von Proteinzuständen, der Identifizierung von Schlüsseldistanzbeschränkungen, der NOE-Netzwerkanalyse und der semiquantitativen Analyse des Proteinkorrelationsnetzwerks.

Umfangreiche Tests der neuen Werkzeuge zur Extraktion und Untersuchung von Protein-korrelierten Bewegungen in Form von Strukturkorrelationen führten uns zu der Entdeckung, dass konventionelle flüssige NMR-Proteinstrukturen im Einzelzustand, die alle zustandsabhängigen Informationen konstruktionsbedingt ausmittelt, gültige strukturelle Korrelationen enthalten. Hier stellen wir einen möglichen Mechanismus zur Beibehaltung solcher strukturellen Korrelationen am Beispiel von Minimalsystemen vor und validieren ihn mit synthetisch aufbereiteten Daten. Wir zeigen auch gültige strukturelle Korrelationen am Beispiel einer experimentellen Flüssig-NMR Einzustands-Struktur des Proteins cyclophilin A. Darüber hinaus präsentieren wir Strukturkorrelationsergebnisse für die gesamte PDB-Datenbank und Indikationen, die darauf hindeuten, dass strukturelle Korrelationen der Einzustands-Flüssig-NMR-Proteinstrukturen sich mit allosterischen Proteinstellen überlappen und Einblicke in die Proteinallosterie geben könnten.

Table of contents

Table of abbreviations

AI	artificial intelligence
ASD	allosteric database
CCR	cross-correlated relaxation
CPMG	Carr-Purcell-Meiboom-Gill
Cryo-EM	cryogenic electron microscopy
eNOE	exact nuclear Overhauser effect
HSQC	heteronuclear single quantum coherence
MD	molecular dynamics
NMA	normal mode analysis
NMR	nuclear magnetic resonance
NOE	nuclear Overhauser effect
NOESY	nuclear Overhauser effect spectroscopy
PCA	principal component analysis
PCS	pseudocontact chemical shifts
PRE	paramagnetic relaxation enhancement
RDC	residual dipolar couplings
RMSD	root mean square deviation
TOCSY	Total correlation spectroscopy (HCCH is the magnetization pathway)
TROSY	transverse relaxation optimized spectroscopy

Chapter 1: Introduction

Protein dynamics

One of the most relevant topics in the field of structural biology is the elucidation of the molecular mechanisms behind enzyme activity, protein target recognition and transduction of biological signals enabling various cellular pathways. Those fascinating topics have one thing in common, they rely on protein fold and dynamics [3]. Thermal protein motion gives rise to the random protein vibrations and non-random conserved correlated protein motion spanning across the scaffold that gives rise to the protein states and equilibrium [4, 5].

A particularly complex example of the correlated protein motion is a ligand-induced synchronized motion between distant sites, termed allostery. Several mechanisms for such motions have been proposed including the dynamic allostery model [3] and population shift model [2]. The dynamic allostery model is based on a statistical thermodynamics model able to quantify allosteric communication in the absence of a conformational change by investigating the effect of ligand binding on thermal fluctuations within a protein. The population shift model is based on ligand-induced structural rearrangements between two distinct protein conformations.

Detailed investigation of the allosteric ligand-protein interaction with the population shift model allows us to further distinguish between the induced-fit model and conformational selection model [6] depending on the time, when the protein rearrangement happens as shown in the Figure 1.1. If the protein rearrangement is happening before the ligand binding, then according to the conformational selection model free protein form exists in equilibrium between two distinct states from which only one is selected by the ligand. If the protein rearrangement is happening after the ligand binding, then according to the induced-fit model the protein allostery can be explained as a difference between free and bound protein conformations.

Despite the high importance of the protein dynamics, investigation of the protein motion at protein conformational states the absolute majority of the protein structures deposited in the protein databank PDB depict proteins in single state resembling a rigid molecule [7].

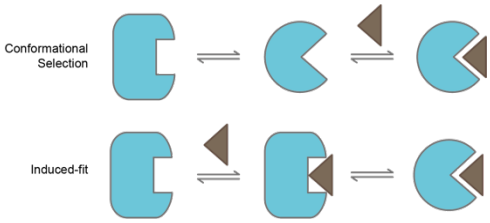

Figure 1.1: Allosteric binding models explaining conserved ligand-induced changes in the protein fold.

Experimental methods probing protein dynamics

Typically, protein allostery is studied for larger systems including enzymes or multimeric proteins with X-ray crystallography or cryogenic electron microscopy (cryo-EM) by quantification of the structural differences between a ligand-free and a ligand-bound protein fold [8]. Extensive studies of allosteric proteins and enzymes allowed to construct an independent allosteric database (ASD) that is primarily populated by the X-ray crystallography and cryo-EM protein structures [9-12].

The use of liquid-state nuclear magnetic resonance (NMR) allows to investigate proteins at their native state giving access to the numerous ensemble-averaged parameters not only for the ligand-free and a ligand-bound proteins, but also for their intermediates. Recent studies were able to show allosteric communication within a protein fold even for proteins with less prominent ligand induced structural changes with NMR [1, 13-15]. However, the use of NMR is limited by the short NMR signal relaxation time of larger proteins and restricts allosteric studies to small protein or individual protein domains below 35 kDa [16].

Multiple techniques that are not based on protein structure were developed to study protein dynamics using NMR as shown in the Figure 1.2. The first type of such techniques is based on NMR relaxation. Protein side-chain rotations can be probed by determination of different relaxation rates including spin-lattice relaxations T_1, spin-spin relaxation T_2 or heteronuclear NOEs [17]. Those model-free relaxation methods provide order parameters (S^2) that quantify the degree of the angular amplitude of the internal motion and correlation times (τ_e) that quantify the local protein dynamics. Both parameters are sensitive to a sub-nanosecond motions that are typically associated with rotation of the residue sidechains [17]. Backbone rearrangements can be probed with T_2 measurements, most notably with Carr-Purcell-Meiboom-Gill (CPMG) methods [18]. The second type of NMR methods probing protein dynamics is based on specific protein labeling including paramagnetic relaxation enhancement (PRE) and pseudocontact chemical shifts (PCS) that can indicate the paramagnetic label environment interactions that are longer than the Nuclear Overhauser Effect (NOE) limit of 5 Å [19-21]. The third type of NMR methods probing protein dynamics is based on the description of the protein-ligand interactions by titration of the ligand and observing protein complex

intermediates with heteronuclear single quantum coherence spectroscopy (HSQC) that is used for the studies of protein allostery as it accurately probes dynamic differences in atomic environment of the protein backbone upon addition of the binding partner and therefore captures ligand-induced protein scaffold rearrangements and allosteric interactions. However, NMR can also be used to produce protein structural ensembles using NOEs that sample protein conformational space and allow to study protein correlated motion.

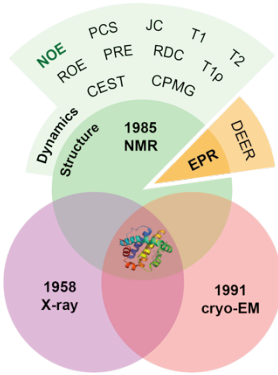

Figure 1.2: Overview of the experimental techniques probing protein structure and dynamics

Paradigm shift for protein NMR

Recently, the introduction of the direct electron detectors and general improvement of the optics and data analysis software for the cryo-EM allowed to significantly enhance the resolution of the experimental protein structures and revolutionized the field of structural biology [22]. Starting by the year 2012 the pace of protein structure elucidations with cryo-EM has been exponentially increasing. At the same time came remarkable advances in the computational prediction of the protein structures with AlphaFold [23]. It allowed to make highly accurate prediction of protein structures. Aforementioned advancements in the field of protein structure determination affect conventional protein structure elucidation methods using NMR. Protein NMR field moves towards protein dynamics field that allows to fully exploit unique advantages of NMR as it can investigate proteins, protein complexes, and their intermediates in their native state with an abundance of the experimental techniques probing protein dynamics from which the eNOE technique is of a particular interest.

Time scales of protein motion

The nature of the protein dynamics and protein motions is ultimately dependent on the time scale of the observation. Protein dynamics can be separated in the fast mode including side chain rotations covering the picosecond and nanosecond time scale and slow mode including protein-ligand interaction, backbone rearrangement, protein folding and unfolding and protein allostery covering second, millisecond and microsecond time scale [24]. Fast protein dynamics on the nanosecond and picosecond time-scale can be probed with NMR relaxation methods including T_1, T_2 and heteronuclear NOE measurements [17]. Slower, millisecond time-scale can be probed with CPMG methods based on the T_2 relaxation [18]. However, this relaxation methods leave a gap at microsecond time-scale that can be observed with PRE NMR measurements [19-21].

Protein motion encoded into the NOESY-derived multi-state NMR structures does not strictly depend on the time-scale as conformers are calculated independently. However, the fast-exchange assumption invoked by the NOESY cross-peak analysis limits the observable motion to be under the low millisecond time-scale. Since protein backbone in NMR ensembles is better resolved than sidechains the major state-dependent observations of protein motion are coming from the backbone rearrangement motions which further limits the protein motion time scale to be above the nanosecond range. Taken together it is possible to predict the most probable time-scale of the visible protein motions from the multi-state NMR protein ensembles to be in the microsecond range.

Exact NOE structure calculations

NMR is a leading technique for experimental studies of dynamics and multi-state structural information of biomolecules because it provides information at atomic resolution and can be measured in aqueous solution. Development of the multidimensional nuclear Overhauser effect spectroscopy (NOESY) allowed to resolve numerous NOEs present in multiple dimensions and use the NOE-based average distances to solve protein structure [19, 25]. Followed by the emergence of the concept of the protein structure solving with liquid NMR an automated software for the protein structure solving CYANA was developed [26].

Recent advances in the field of protein NMR allow us to gain insights into the protein motion at atomic resolution by determining multiple protein states using NMR supplied with a plethora of experimental restraints including residual dipolar couplings (RDC), cross-correlated relaxation (CCR), paramagnetic relaxation enhancement (PRE) and NOE restraints [19, 26-32]. Remarkable advancement in the field of the NOE-based protein structure determination including eNOEs can yield experimental time-averaged ^{1}H-^{1}H distances with the resolution of up to 0.1 Å that together with the structure calculation of multiple protein states, correction of spin diffusion with eNORA and an automated implementation in the eNORA2 package within CYANA allows a straight forward execution of eNOE based structure calculations [33-35] yielding multiple protein states at atomic resolution. So far eNOEs were successfully applied to WW domain, protein GB3, cyclophilin A, and protein ubiquitin [35-38].

Multi-state structure calculations

Distance restraints that are obtained during the analysis of the cross-peaks from NOESY spectrum provide the ensemble-averaged distances. However, if the protein is in exchange between multiple conformations the average distances cannot be simultaneously valid for a single protein state. Therefore, assuming multiple coexisting protein conformations a simultaneous optimization of multiple protein states is performed with the software CYANA. The multi-state structure calculation is based on the minimization of the target function calculated by comparing the simulated distance back calculated from all optimized protein states with experimental upper and lower limit restraints extracted from the NOE cross-peak. Individual protein states are kept in proximity of each other by symmetry restraints, but local movements with amplitude below 1.2 Å are allowed. Such approach provides an additional flexibility to the protein fold and allows to observe concerted protein motion.

Additional degrees of freedom provided by introduction of multiple protein states allow to minimize the overall violation of the experimental restraints and therefore the value of the CYANA target function. On the local level the network of the NOE distance restraints minimized for multiple states allows individual residues or local protein features to split between local minima represented by distinct combinations of residue dihedral angles or relative position of a certain local feature of the protein fold. On the global level, collective state-specific minima give rise to protein correlated motion that carries important biological information.

Spin diffusion correction

Spin diffusion is a major factor causing inaccuracy in deriving distances from the NOEs [39]. Spin diffusion can be theoretically corrected if all NOE cross and diagonal peaks can be measured unambiguously, which is unrealistic given a limit to NMR sensitivity [40]. Therefore, an alternative version of the spin diffusion correction called exact NOE by Relaxation matrix Analysis (eNORA) was implemented as a part of the eNOE technique that relies on the given 3D protein structure [33]. Furthermore, spin diffusion effects are suppressed by shorter NOE mixing times.

The full-relaxation matrix approach can be used to correct magnetization buildups for spin diffusion [33]. In this approach, we calculate the NOE magnetization transfer on the protein scaffold with and without spin-diffusion. Then, for each cross-peak and for each mixing time we will calculate the correction factor as a ratio between two theoretical cross-peak intensities and apply it to the experimental data.

The time evolution of the NOESY intensities $I(\tau_{mix})$ can be described by the multispin Solomon equations [34, 41, 42] as follows:

$$I(\tau_{mix}) = I(0)e^{-\boldsymbol{R}\tau_{mix}},$$

where $\boldsymbol{R}$ is a relaxation matrix containing auto and cross-relaxation rate constants ρ_i and σ_{ij}.

$$\boldsymbol{R} = \begin{pmatrix} \rho_1 & \cdots & \sigma_{1N} \\ \vdots & \ddots & \vdots \\ \sigma_{N1} & \cdots & \rho_N \end{pmatrix}$$

Under assumption of ideal two-spin system it is possible to fit the intensity buildup with non-linear isolated spin-pair approach (ISPA) [33]. Intensity dependence on the mixing time of the NOESY cross-peak can be described as following:

$$\frac{I_{ij}(t)}{I_{ij}(0)} = \frac{I_{ji}(t)}{I_{ji}(0)} = \frac{-\sigma_{ij}}{\lambda_+ - \lambda_-}[e^{-\lambda_- t} - e^{-\lambda_+ t}]$$

with

$$\lambda_\pm = \frac{\rho_i + \rho_j}{2} \pm \sqrt{\left(\frac{\rho_i - \rho_j}{2}\right)^2 + \sigma_{ij}^2}$$

As an approximation we assume that NOESY diagonal peaks decay as single-exponential functions:

$$\frac{I_{ii}(t)}{I_{ii}(0)} = e^{-\rho_i t}$$

In context of the aforementioned formalism, it is possible to fit individual auto-relaxation rates ρ_i from the experimental decays of the NOESY diagonals and use them to fit all apparent cross-relaxation rates σ_{ij} with an ISPA approach from the experimental cross-peak buildups. Cross-relaxation rates σ_{ij} acquired with this approach does not negate the influence of spin diffusion.

In order to calculate spin diffusion correction factors with an eNORA2 simulation previously determined protein structure is used. For this a theoretical relaxation matrix is populated with theoretical NOE cross-relaxation values:

$$\boldsymbol{R}^* = \begin{pmatrix} \rho_1 & \cdots & \sigma_{1N}^* \\ \vdots & \ddots & \vdots \\ \sigma_{N1}^* & \cdots & \rho_N \end{pmatrix},$$

where the theoretical cross relaxation rates are calculated from the distances between spins and rotational correlation time:

$$\sigma_{ij}^* = \left(\frac{\mu_0}{4\pi}\right)^2 \frac{\gamma^4 \hbar^2}{10} \frac{\tau_c}{r_{ij}^6} \left[\frac{6}{1 + 4\omega_0^2 \tau_c^2} - 1\right]$$

This allows to simulate theoretical magnetization buildups through all possible magnetization pathways:

$$I^*(\tau_{mix}) = I(0) e^{-\boldsymbol{R}^* \tau_{mix}}$$

and compare resulting cross-peak intensities to those extracted under the assumption of the isolated two spin system from the theoretical auto and cross-relaxation rates:

$$I(\tau_{mix}) = I(0) \frac{-\sigma_{ij}^*}{\lambda_+^* - \lambda_-^*} [e^{-\lambda_-^* \tau_{mix}} - e^{-\lambda_+^* \tau_{mix}}]$$

to correct experimental cross-peak intensities for spin diffusion:

$$I_{cor}^{exp}(\tau_{mix}) = I^{exp}(\tau_{mix}) F(\tau_{mix}),$$

where the correction factor is:

$$F(\tau_{mix}) = \frac{I(\tau_{mix})}{I^*(\tau_{mix})}$$

In some cases, measurement of multiple NOESY spectra is not practical. In cases where only a single NOESY spectrum is available it is not possible to fit exponential decays for NOESY diagonal peaks and buildups for the NOESY cross-peaks as it is done in eNORA2 for the spin diffusion correction [34]. However, methods for the spin-diffusion correction from a single NOESY spectrum are available [43, 44] and the Riek group actively investigates alternative spin diffusion correction algorithms based on a single NOESY spectrum.

Machine learning application in protein NMR

The adoption of the eNOE approach by the NMR community does not happen with a rapid pace presumably due to the long NMR acquisition time necessary to acquire multiple 3D-NOESY spectra and high required spectrum quality as is necessary to resolve large number of cross-peaks at relatively low NOE mixing times and generate large number of the distance restrains that would overdetermine the NOE network and allow for the resolution of multiple states. Furthermore, demanding and highly specialized computational procedure are required to calculate multiple protein states. However, the Riek group attempts to automatize multiple steps of the demanding multi-state NMR structure calculation with help of artificial intelligence (AI) and machine learning. Recent advances hint that in the near future it might be possible to solve an NMR protein structure with a single click and in a fully autonomous fashion by the application of the automated and spectrometer-integrated protein structure calculation software. Those advances are based on the application of the artificial intelligence for the peak picking from the various multi-dimensional spectra, automated protein assignment using FLYA and protein structure calculation using CYANA [31, 45].

Qualitative description of the protein correlated motion

There are two levels of complexity associated with multi-state NMR protein structure elucidation. First, the ensemble of the protein coordinates should be calculated according to the exact NOE approach and second, meaningful biological information should be extracted from the ensemble. So far, protein state interpretation was performed manually by repeated selection of key residues, separation of conformers into states according to the Ramachandran statistics of the selected residue or according to some local features of the protein ensemble that allow to separate protein conformers in equally populated states and observing to which protein sites this conformer separation spreads without randomization along the aligned 3D structures of the protein ensemble. With this method correlated motion was successfully evaluated and reported for all previous eNOE structures including WW domain [36], protein GB3 [35] and cyclophilin A [37]. However, this method is based on the demanding and subjective evaluation of small differences between calculated multi-state conformers. Therefore, a novel objective and automated method for the extraction of the correlated motion from the protein ensembles, PDBcor, was introduced [46].

Chapter 2: PDBcor: An Automated Correlation Extraction Calculator for Multi-State Protein Structures

This chapter is an adaptation from the following manuscript: Ashkinadze, Dzmitry, et al. "PDBcor: An automated correlation extraction calculator for multi-state protein structures." *Structure* (2021).

Author's contribution: D.A. developed PDBcor P.K. improved the machine learning and visualization aspects of the PDBcor and implemented PDBcor server H.K. proposed to validate PDBcor on previous eNOE structures P.G. contributed significance thresholding P.G. and R.R. supervised the project D.A., P.G. and R. R. wrote the manuscript. All authors discussed the results and contributed to the final manuscript

Introduction

Protein motion including correlated motion can be extracted from protein structural ensembles generated under assumption of multiple protein states. Multi-state protein structures are typically determined by experimental methods including NMR using previously mentioned eNOE approach [27, 28, 33, 47], by different class selections in cryo-EM-derived structure determination [48], or by the presence of distinct X-ray structures due to different crystal packings or the same crystals exposed to a strong electric field [49]. Alternatively, such protein ensemble structures could be generated with molecular dynamics (MD) canonical ensemble simulations in presence or absence of experimental data [50-52]. Conventionally, correlated motion is extracted in the form of residue-based cross-correlation matrices from MD trajectories [53-56] or alternatively from the superimposed structural ensembles either with principal component analysis (PCA) [55, 57] or normal mode analysis (NMA) [58] based approaches. According to the PCA-based or NMA-based approach extracted residue cross-correlation values are calculated as covariances between residue coordinate vectors. However, the use of the absolute Cartesian coordinates requires the conformer superposition that is impossible to do objectively if multiple protein motions are present.

In this work we present a method that does not require any structure superposition and therefore is unbiased due to the fact that it is based solely on distance and angle statistics of individual structural entities. PDBcor performs an objective and automated correlation analysis of multi-state protein structures, which can be used for the elucidation of biologically important correlated motion. With the help of information theory, it is possible to extract residue-based protein correlations in fully automated fashion. Information about such biologically relevant correlations is vital for our understanding of proteins. PDBcor is publicly available as a Python executable at https://github.com/dzmitryashkinadze/PDBcor or as a web server at http://pdbcor.ethz.ch.

Theory

The workflow of the correlation extraction procedure with PDBcor is shown in Figure 2.1. First, an input structure bundle is subjected to significance thresholding that filters out spurious correlations. Second, interresidual distances are used to cluster conformers. Finally, residue clusterings are compared to obtain a correlation matrix.

Objective extraction of correlated motion

PDBcor relies on structure comparison based on a statistical analysis of interresidual distances or dihedral angles within individual conformers that does not require any superpositions. Conventionally a superimposed ensemble of protein conformations is visually sorted based on certain local protein features. For example, if protein conformers are sorted according to the relative position of a particular α-helix, neighboring regions might be sorted correctly and therefore correlate to the α-helix, but such sorting is typically not coherent throughout the whole protein scaffold [14]. In order to systematically study those correlations an ensemble of multistate protein conformations is repeatedly clustered for each residue with the aim to extract correlations between protein residues. Residue correlations are evaluated by computing a similarity between two arbitrary conformer clusterings.

Significance thresholding

Correlations extracted with PDBCor are based exclusively on similarity between residue clusterings (see below). As such, they are largely independent of the degree of separation between states. In some well-defined structural bundles individual states might therefore be identified that are closer to each other than the amplitudes of random thermal motion. This might lead to spurious distance correlations. To avoid such artifacts, a small amount of Gaussian noise is added to the atomic coordinates:

$$r'^{(j)}_{im} = r^{(j)}_{im} + \delta^{(j)}_{im}$$

where $r^{(j)}_{im}$ is the position of atom *m* in residue *i* of conformer *j*, which is obtained with Biopython [59], and $\delta^{(j)}_{im}$ is a vector of three independent, normally distributed random numbers with zero mean and standard deviation *σ*. This leads to random mixing of insignificantly separated protein states and suppresses spurious distance correlations.

The noise amplitude *σ* should be set such that it is sufficient to remove background correlations with amplitudes below that of thermal motions and experimental uncertainties but does not exceed the separation between significantly different protein states that would remove correlations of interest. A standard value of 0.5 Å was used for all presented experiments as a value that resembles the fast (ps) order parameter of 0.8 that has been measured in proteins by NMR [60]. However, PDBcor allows also to switch off the noise generator completely.

Residue-based conformer clustering

For the purpose of clustering, each residue *i* is represented by a single point, given by its centroid coordinates in conformer *j*

$$x^{(j)}_i = \frac{1}{M_i} \sum_{m=1}^{M_i} r'^{(j)}_{im}$$

where $\boldsymbol{M_i}$ is the number of atoms of residue i that are considered for the correlation calculation. The scope of input atoms can be predefined to be either the backbone atoms, the sidechain atoms, or all atoms of the residue (see below). From the centroid coordinates, we construct a distance matrix D with elements

$$D^{(j)}_{ik} = |x^{(j)}_i - x^{(j)}_k|$$

Each row of the distance matrix contains the distances between the center of a given residue *i* and the centers of the other residues *k* and thus defines the relative location of

the residue that can be used as a fingerprint of a given conformer. In the case of N distinct residue-based protein conformations, we expect that interresidual distances of all conformers from a given structure ensemble can be grouped into N clusters. Using this assumption, conformers are clustered based on their interresidual distances into N groups for each residue using the Gaussian Mixture Model (GMM) algorithms [17]. This yields, for each residue *i*, a distance clustering vector c_i with elements $c_{ij} \in \{1, \dots, N\}$ that stores the cluster labels of all conformers $j = 1, \dots, N$. The total set of protein interresidual distances that is used as input to the PDBcor is highly redundant as number of distances is proportional to the number of residues squared. However, conformers are clustered independently for each residue and for a selected residue a non-redundant set of distances from the selected residue to the rest of the protein is used.

As an alternative to distance-based clustering, the clustering can also be based on the backbone ϕ, ψ, ω and side chain $\chi_1, \chi_2, \chi_3, \chi_4, \chi_5$ torsion angles. For residues with less than five side chain torsion angles, the undefined χ values are set to zero. As in the distance case an angular matrix $\Phi^{(j)}$ is formed by the eight dihedral angle values of each residue in the conformers $j = 1, \dots, N$. It is used to cluster conformers into *N* groups using GMM. In complete analogy to the distance-based case, this yields, for each residue *i*, an angular clustering vector c_i^a with elements $c_{ij}^a \in \{1, \dots, N\}$ that stores the cluster labels of all conformers $j = 1, \dots, N$.

Evaluation of correlated motion

Correlation extraction from the clustering matrix is possible using information theory [18-20]. Two arbitrary clustering results are represented by two discrete variable vectors X and Y. One of the most extensively studied measures specifying the amount of correlation between two discrete variable vectors is the mutual information $I(X, Y)$ [21]:

$$I(X,Y) = \sum_{x,y=1}^{N} \quad p(x,y) log \frac{p(x,y)}{p(x)p(y)},$$

where x and y are cluster labels of clusterings X and Y with probabilities $p(x) = p(X = x)$, $p(y) = p(Y = y)$ and joint probability $p(x, y) = p(X = x, Y = y)$. The mutual information

tells us how much the conformer clustering of one residue tells us about the conformer clustering of another residue. A variant of the mutual information that was specifically developed for clustering comparison is the adjusted mutual information $I^{\star}(X,Y)$ [22]:

$$I^{\star}(X,Y) = \frac{I(X,Y) - E\{I(X',Y')\}}{max\{H(X),H(Y)\} - E\{I(X',Y')\}}$$

where $E\{I(X',Y')\}$ is the expected value of the mutual information for an ensemble of random, uncorrelated vectors X' and Y', and $H(X)$ is the entropy of the variable X:

$$H(X) = -\sum_{x} \quad p(x)\, log\, p(x)\ ,$$

where $p(x)$ is the probability of cluster x. Note that $I^{\star}(X,Y) = I^{\star}(Y,X)$ for any pair of clusterings, and $I^{\star}(X,Y) \approx 0$ between two random clusterings. The adjusted mutual information yields a correctly normalized value measured in bits that is a suitable measure for the correlation between protein residues.

Given a clustering matrix (C^{α} or C^{d}), all residue pair combinations are compared using the adjusted mutual information, describing similarity between residues. The adjusted mutual information scores for residues *i* and *j* form a symmetric correlation matrix A with elements $A_{ij} = I^{\star}(c_i, c_j)$ for distance-based clustering, or $A^{a}_{ij} = I^{\star}(c^{a}_{i}, c^{a}_{j})$ for torsion angle-based clustering. Visual inspection of the correlation matrix heatmap (Figure 2.1) provides information about residues or subdomains that are involved in a correlated motion. In addition, the mean value of the elements of the matrix A yields an overall correlation parameter for the structure ensemble.

Both distance and angular correlation analyses are able to detect correlated motion. Nevertheless, distance correlation extraction is more sensitive to the protein motion.

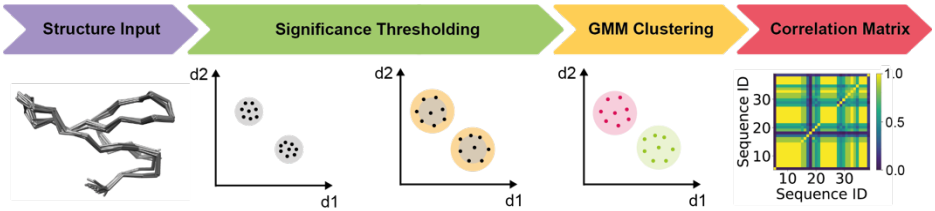

Figure 2.1 Overview of the correlation extraction procedure. First, an input structure bundle (PDB ID 6SVC [23]) is subjected to the noise generator that filters out spurious insignificant correlations. Here an illustrative example is depicted, where conformers existing in two states are shown as points in a scatter plot of two arbitrary distances (for example first is a distance between residues X and Y and second is a distance between residues X and Z). During significance thresholding random displacement of atoms broadens the edges of states so that states separated by less than the amplitude of the noise loose separation. Then, interresidual distances are used to cluster conformers for each residue with GMM (in this case it would be residue X). Finally, a pairwise comparison of the resulting clustering vectors based on their mutual information yields an interpretable correlation matrix.

Global conformer clustering

For visualization purposes, it is useful to get an optimal global (rather than residue-specific) clustering of conformers that can be used for highlighting state-specific features in a protein ensemble superposition view. For example, the two sets of clustered conformers within a two-state structure ensemble can then be colored differently as shown in the Figure 2.2 below.

To this end, we cluster the conformers according to the clustering c_i of the residue *i* that has the highest average correlation to the other residues of the protein. Since the protein ensemble superposition is made according to the protein coordinates, the distance correlation matrix A is used to calculate the average residue correlations.

Versatility of PDBcor for backbone and sidechain correlations

The correlation extraction procedure allows to control the protein region from which correlations are extracted by filtering the input data. In particular, backbone correlations can be extracted by utilizing only backbone atom coordinates and backbone dihedral angles. Similarly, sidechain or total (backbone and sidechain) correlations can be extracted. This possibility might be in particular interesting for some experimental methods including NMR for which the backbone structure is better resolved than side

chains. Therefore, extraction of backbone correlations could be beneficial for the resolution and sensitivity of protein correlations.

Results

Spatial correlations in protein structures

Three different protein ensembles from the Protein Data Bank that have been determined by liquid-state NMR act as examples for a non-correlated protein ensemble (Figure 2.2a), a locally correlated protein ensemble (Figure 2.2b), and a globally correlated protein ensemble (Figure 2.2c). The structure bundles were analyzed by PDBcor with the assumption that an ensemble of structures samples the conformational space of a protein with residue-based two-state dynamics, regardless of the structure origin.

Distance correlation matrix heatmaps of non-correlated systems do not show any significant correlations (visualized by yellow spots in the heatmap) (Figure 2.2d). Optimally clustered conformers of non-correlated systems are typically non balanced with one state dominating the other one. The most probable explanation for the absence of correlations in such structure ensembles is a violation of the two-state model assumption.

As opposed to non-correlated systems, distance correlation matrix heatmaps of locally correlated systems show correlations that are localized to distinct regions of the protein structure. Optimally clustered conformers of locally correlated systems can be visually separated into two states in their corresponding protein correlation site. Correlation lights up as yellow spots in the heatmap (Figure 2.2e). This correlation between α-helix 2 (residues 42–51) and α-helix 3 (residues 70–78) can also be seen in the structure superposition and coloring according to the global conformer clustering (Figure 2.2b).

Conformers from globally correlated protein ensembles can be unambiguously separated. It can be easily visually confirmed as protein states do not overlap well due to significant differences between protein states (Figure 2.2c). Since a global separation does not depend on the choice of the residue, there are pairwise correlations between most residues and consequently most of the distance correlation heatmap turns yellow (Figure 2.2f).

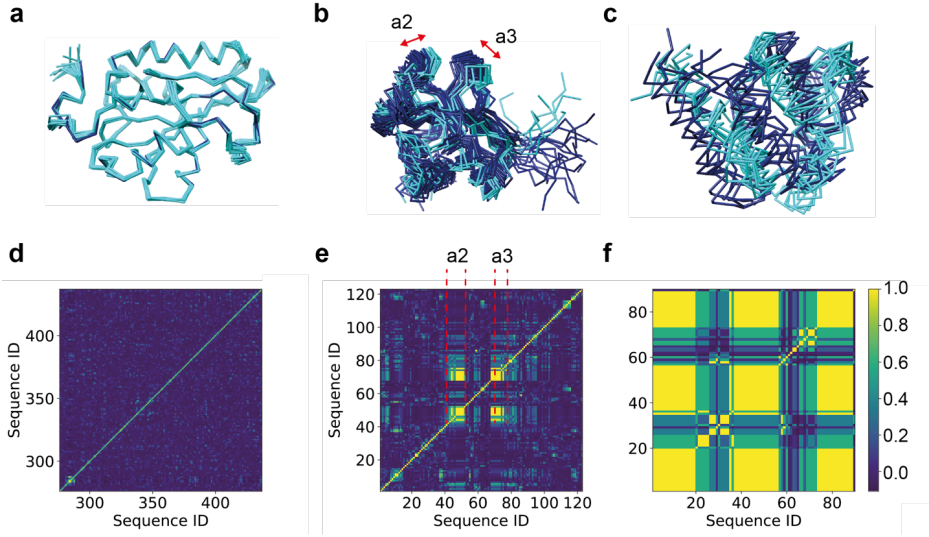

Figure 2.2 Distance correlation matrix heatmaps (bottom panel) and optimally clustered bundles of proteins (top panel) sorted in ascending order of structural correlations. (a) Solution structure of the C-terminal domain of the human eEF1Bγ subunit (PDB ID: 1PBU) is depicted as an example of a non-correlated protein system. The distance correlation matrix heatmap does not show any significant correlations (yellow spots) and a single state (cyan) dominates among the optimally clustered conformers. (b) Solution structure of the Sma0114 (PDB ID: 2LPM) is depicted as an example of a locally correlated system. The distance correlation matrix heatmaps shows correlations that are localized to α-helices 2 and 3, whereas its optimally clustered conformers correlate also only in the regions of α2 and α3. (c) Solution structure of the PEA-15 Death Effector Domain in Complex with ERK2 (PDB ID: 6P6C) is depicted as an example of globally correlated system. Its distance correlation matrix heatmap is mostly correlated with exception of few uncorrelated regions (visible as blue stripes) and its conformers are unambiguously separable. The conformer separation can be easily visually confirmed due to significant differences between protein states.

Correlations of WW domain, protein GB3 and cyclophilin

PDBcor was benchmarked on three model systems: WW domain of PIN1 (Figure 2.3a; PDB ID 6SVC; [23]), the protein GB3 (Figure 2.3b; PDB ID 2LUM; [24]) and cyclophilin A (Figure 2.3c; PDB ID 2MZU; [25]). For all three systems multi-state structure ensembles were determined by solution state NMR based on eNOEs [24]. The detailed time-intensive study of the multi-state structures using subjective superpositions of conformers and objective angular correlations yielded the presence of correlated motion at atomic resolution in all three systems [23-25].

The automated evaluation of the WW domain with PDBcor identifies a globally correlated network (Figure 2.3d). This shows that experimental restraints were able to separate two WW states.

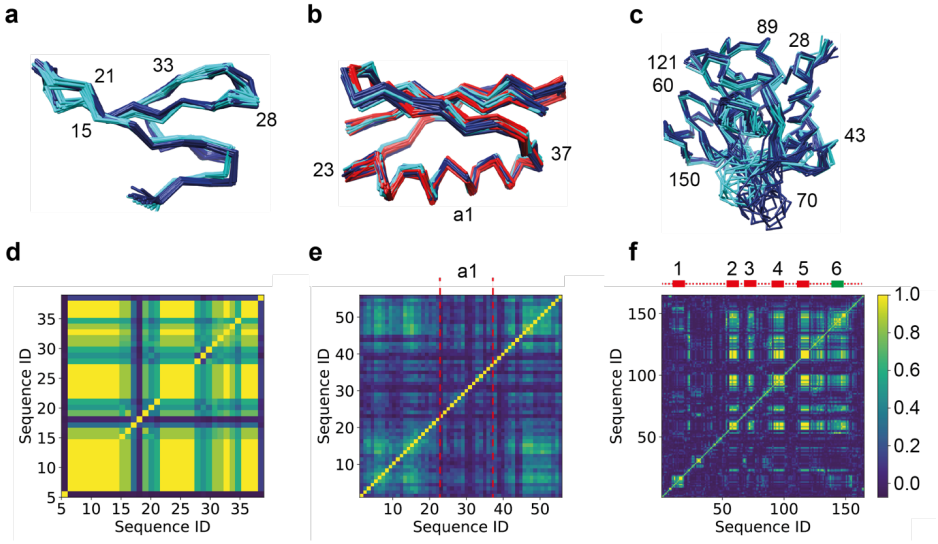

Figure 2.3 Automated correlation extraction results for the WW domain of PIN1 (a, d; PDB ID 6SVC; [23]), protein GB3 (b, e; PDB ID 2LUM; [24]) and cyclophilin A (c, f; PDB ID 2MZU; [25]). The top panels (a, b, c) illustrate the superimposed bundles of conformers, colored according to the optimal global distance-based clustering. The bottom panels (d, e, f) illustrate the backbone distance correlation matrix heatmaps. For the WW domain, the optimally colored backbone bundle (a) and its distance correlation matrix heatmap (d) both identify a globally correlation network. The distance correlation matrix heatmap of GB3 (e) identifies a system that is weakly correlated everywhere except a region covering the α-helix (residues 23–37) and its neighboring residues, highlighted with a pair of red dashed lines, as it was reported previously [24]. The backbone distance correlation matrix heatmap for cyclophilin (f) confirms five previously reported correlation sites, including site 1 (residues 9–16), site 2 (54–57), site 3 (64–78), site 4 (101–107), and site 5 (118–127) highlighted in red [25]. Additionally, PDBcor identifies a previously undetected correlation site 6 (137–155), highlighted in green.

The automated evaluation of the protein GB3 with PDBcor reveals a system that is (weakly) correlated everywhere except for the α-helix of residues 23–37 (Figure 2.3e). This finding confirms the previously reported observation of correlated motion across the β-sheet and a lack of correlated motion between the β-sheet and the α-helix [24]. It is noted that the GB3 protein is reported to comprise three states which was successfully analyzed with PDBcor as it generalizes to an arbitrary number of conformational states.

As an example of a larger system, the protein cyclophilin A was evaluated. According to the distance correlation matrix heatmap (Figure 2.3f) five previously reported correlations in regions 1 (residues 9–16), 2 (54–57), 3 (64–78), 4 (101–107), and 5 (118–127) were confirmed [25]. PDBcor did not only find all reported correlation sites, but also an extension of the correlation system to an additional region in the protein, site 6 (residues 137-155). Notably, sites 2–6 form a fully connected correlation network, whereas site 1 correlates only to site 6. In the case of cyclophilin A the strength of PDBcor is apparent: First, it elucidates all statistically significant structural correlations, yielding an extension of the correlation network that had been found manually. Second, in contrast to a tiresome selection by manual inspection, it is fully automated, objective and reproducible.

Conclusions and Outlook

PDBcor can be used to get an optimal conformer separation for further analysis of protein states. Alternatively, further interpretation of PDBcor correlation matrices allows to quantify correlations, identify which part of the protein is involved in correlated motion and pinpoint most prominent correlations between protein sites. Careful examination of the correlation matrix may provide an information about the localization of correlated subsystems for a given protein.

PDBcor correlation amplitude can be interpreted as an information flow between residue pairs. Therefore, PDBcor is not only able to localize the correlation of interest, but also to quantify it. Strong correlation of a residue pair as in Figure 2.2f means that by knowing the state of the first residue we know the state of the second residue. Weak correlation of a residue pair as in Figure 2.3e means that by knowing the state of the first residue we can predict with some certainty the state of the second residue.

Any protein structure ensemble can be analyzed with PDBcor. Nevertheless, meaningful correlations can only be extracted from structure bundles that have been generated with the aim to incorporate information about multiple protein states. Furthermore, a cautious use is indicated for proteins with disordered regions.

The number of protein ensemble structures grows together with a rapid advancement in the field of structural biology [26]. A fraction of such deposited ensemble structures contains the information about correlated motion. The knowledge about such protein correlations is vital for the understanding of protein mechanism of action and should be systematically studied.

Supplementary Information

Application of the PDBcor to MD trajectories

In order to illustrate that PDBcor-based analysis can be applied to the protein structural ensembles originated from techniques other than NMR we analyzed a series of molecular dynamics (MD) trajectories. MD trajectories were downloaded from the MoDEL (Molecular Dynamics Extended Library) [61]. Compressed backbone MD trajectories for WW domain, protein GB3 and cyclophilin A, each consisting out of 10000 frames simulating 10ns, 10ns and 80.5ns were downloaded as PDB IDs 1i6c, 2igd and 2cpl. They were uncompressed with PCAsuite [62], sliced down to 100 conformations with MDTraj [63] and inputted to PDBcor. Resulting structural correlations are summarized in Figure S2.1.

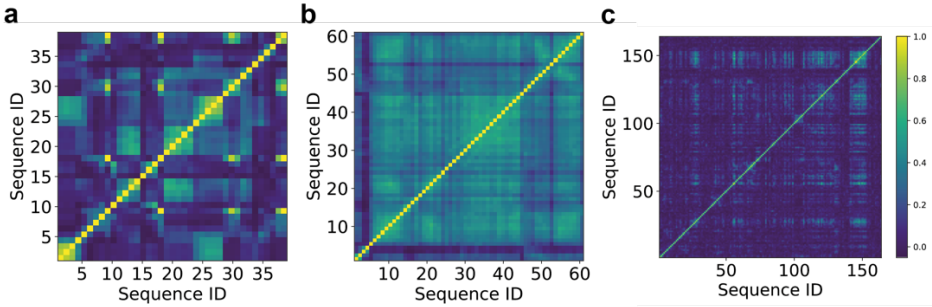

Figure S2.1 Structural correlation analysis with PDBcor of MD trajectories for the WW domain (a), protein GB3 (b) and cyclophilin A (c).

Comparison of the PDBcor to the PCA and NMA-based methods

In order to illustrate high sensitivity of the PDBcor we compared it to the conventional PCA-based technique THESEUS [57] and NMA-based technique WEBnm@ [58]. THESEUS performs structure alignment with maximum likelihood algorithm followed by PCA of the aligned protein coordinates that optimizes a correlation matrix. Unlike PDBcor, PCA-based approaches require structure superposition and therefore they are biased to the way superposition was done. Furthermore, PCA-based approaches calculate correlations from atomic deviations from the mean structure that are deduced from atom coordinates, whereas in PDBcor no assumption of the mean structure is made and interresidual distances that are more sensitive to the less pronounced, but statistically significant protein rearrangements are used. In turn NMA-based approaches are based on analysis of torsion angles, whereas PDBcor is based largely on the interresidual distances and therefore PDBcor by design is more sensitive to the integrated correlated motion of secondary structure elements or protein domains.

Structural correlation of the cyclophilin A, a known and reported allosteric molecule were analyzed with PDBcor, Thesaurus and WEBnm@ and compared in Figure S2.2. Whereas PDBcor results overlap with reported findings as discussed in Figure 2.3, THESEUS and WEBnm@ techniques failed to reproduce them.

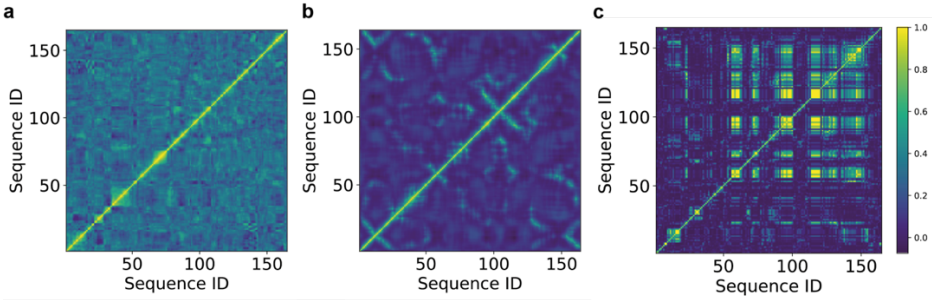

Figure S2.2 Structural correlation analysis of cyclophilin A with THESEUS (a), WEBnm@ (b) and PDBcor (c).

Chapter 3: Atomic resolution Protein Allostery from the multi-state Structure of a PDZ domain

This chapter is an adaptation from the manuscript in preparation: Dzmitry Ashkinadze, Harindranath Kadavath, Celestine Chi, Michael Friedmann, Dean Strotz, Pratibha Kumari, Martina Minges, Riccardo Cadalbert, Stefan Koenigl, Peter Güntert, Beat Vögeli*, Roland Riek*, Atomic resolution Protein Allostery from the multi-state Structure of a PDZ Domain

Author's contribution: D.A. refined and solved free eNOE PDZ2 structure, assigned and solved ligand-bound eNOE PDZ2 structure C.C. and R.C. prepared PDZ2 samples, B.V. and H.K. measured NMR spectra, M.M. and R.R. assigned and solved free PDZ2 structure B.V. and R.R. supervised the project D.A., H. K., B.V. and R.R wrote the manuscript. All authors discussed the results and contributed to the final manuscript.

Introduction

One of the most studied group of allosteric molecules are PDZ domains. The family of PDZ domains is crucial for protein-protein recognition and protein complex assemblies in multicellular organisms [64]. PDZ domains recognize carboxyl-terminus of various target proteins and take part in many cellular processes including cell growth and proliferation. Second PDZ (PDZ2) domain displays a compact fold out of six β-strands, two α-helices and a unique flexible loop at the bottom of the binding pocket [65]. The strands of the protein form an antiparallel β-sheet that serves as a platform for target molecule binding. Protein human tyrosine phosphatase 1E (hPTP1E) contains a PDZ2 domain and mediates a series of crucial biological processes such as protein-protein interaction [66, 67], signaling [68] and apoptosis [69]. Solution NMR structures of the PDZ2 domain of hPTP1E were solved for a free form as well as for the form bound to the C-terminal peptide derived from the Ras-associated guanine nucleotide exchange factor 2 (RA-GEF2) [70, 71]. PDZ2 domain of hPTP1E binds RA-GEF2 by a β-strand addition between strand β2 and α-helix 2 similar to other PDZ domains [72]. PDZ2 allostery was studied with various techniques including the use of evolutionary data [2] and protein dynamics data [1]. Both approaches underly the importance of residues Ile20, Val85 and Val61 for the protein allosteric network.

In this study we use the eNOE approach to investigate multi-state structures of both free and bound forms of the PDZ2 domain of hPTP1E with aim to elucidate its correlated motion, ligand binding mode, protein allostery and correlated motion by investigating individual protein states at atomic resolution as enabled by eNOE approach.

Results

Ligand-induced dynamic changes of PDZ2 domain

Heteronuclear 2D NMR spectroscopy was applied in order to gain a qualitative understanding of the PDZ2 domain of hPTP1E binding to the C-terminal peptide derived from the Ras-associated guanine nucleotide exchange factor 2 (RA-GEF2; Ac-ENEQVSAV-COOH) and allosteric interactions. A [^{1}H,^{15}N]-HSQC spectrum was acquired at 298 K for a uniformly enriched ^{15}N-labeled PDZ2 domain both for a free form and bound to the peptide supplied in a two-fold excess. An overlay of the [^{1}H,^{15}N]-HSQC spectra together with a chemical shift perturbation (CSP) map are shown in Figure 3.1. Averaged atom-weighted chemical shift perturbations (nitrogen shifts were taken with weight of 0.3) were mapped on the later calculated apo protein structure to summarize a scope of the allosteric system of the PDZ2 domain. Absolute CSP values are shown in the Figure S3.1. In line with previous reports residues of the PDZ2 binding site showed significant chemical shift perturbations together with a number of allosteric residues as expected from a highly dynamic scaffold [1, 65]. In details, significant CSPs are observed for residues at the β-strand 2 and α-helix 2 which sandwich the ligand upon binding with less prominent CSPs in the flexible loop Gly24-Gly34, β-strand 3, and α-helix 1 of which the latter two are far away from the binding site and thus have been identified as allosteric sides [1].

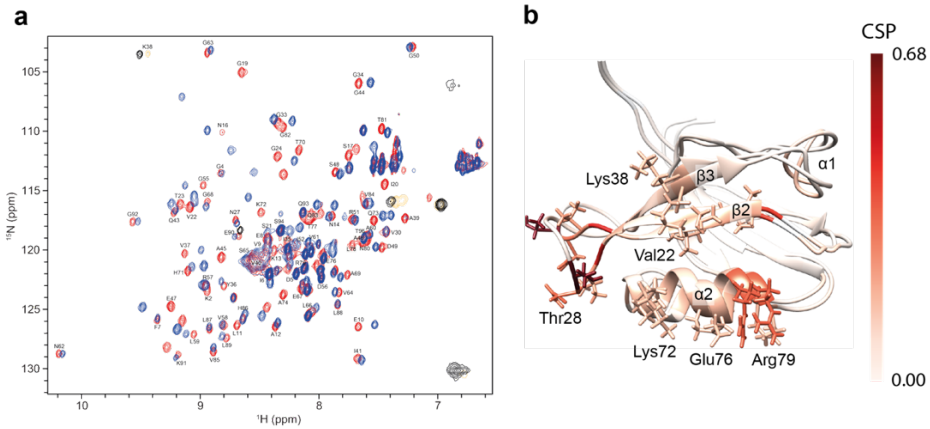

Figure 3.1: Ligand-binding induced conformational changes measured by chemical shifts. [^{1}H,^{15}N]-HSQC spectra for the PDZ2 domain in apo form (red) and bound to the RA-GEF2 peptide (blue) (a). Residues of the two PDZ2 apo states with lowest CYANA target function were colored according to the geometric mean of the chemical shift perturbation of ^{15}N and ^{1}H (with nitrogen shifts taken with weight of 0.3) (CSPs) as indicated by the bar on the right (b).

Multi-state structure determination of the PDZ2 domain

For apo PDZ2 experimental restraint collection included 1553 distance restraints from eNOEs extracted from a set of 3D [^{15}N,^{13}C]-resolved [^{1}H,^{1}H]-NOESY-HSQC spectra at 8, 16, 24, 32, 40, 50 and 80ms NOESY mixing times (with 410 bidirectional distance restrains with highest precision of 0.1 Å). In addition, 65 scalar couplings were collected that resulted in ~17 restraints per residue as summarized in Table S3.1. Similarly for the complex PDZ2 1484 distance restraints from eNOEs and 65 scalar couplings that resulted in ~16 restraints per residue were collected as summarized in Table S3.2.

In general, with this experimental input multi-state protein calculation with the program CYANA can be performed. CYANA simultaneously optimizes multiple protein states by minimization of the target function calculated by comparing the simulated distance back calculated from all optimized protein states with experimental upper and lower limit restraints extracted from the NOE cross-peak. Individual protein states are kept in proximity of each other by symmetry restraints, but local movements with amplitude below 1.2 Å are allowed. Such approach provides an additional flexibility to the protein fold and allows to observe concerted protein motion. Spin diffusion is a major

factor causing inaccuracy in deriving distances from the NOEs [39]. Spin diffusion can be theoretically corrected if all NOE cross and diagonal peaks can be measured unambiguously, which is unrealistic given a limit to NMR sensitivity [40]. Therefore, an alternative version of the spin diffusion correction called exact NOE by Relaxation matrix Analysis (eNORA) was implemented as a part of the eNOE technique that relies on the given 3D protein structure [33]. So far eNOEs were successfully applied to WW domain, protein GB3, cyclophilin A and protein ubiquitin [35-38]. Furthermore, a program PDBcor was developed to elucidate correlated motion in form of structural correlations in an unbiased and automated way from the distance statistics of individual structural entities in a multi-state structure [46]. It can quantify correlations in structural ensembles, uncover the protein regions that undergo synchronized motion, give insights into the biologically important correlated motion and optimally separate conformers into states. PDBcor uses information theory and systematic clustering of protein conformers with aim to extract mutual information between individual residues [46].

In the case of the PDZ2 domain both apo and holo multi-state protein structures were calculated following the established protocol introduced above [35-37] using eNORA2 for the spin diffusion correction [33, 34] and CYANA for the protein structure calculation with minor modifications [31]. Symmetry restraints that keep structural entities in proximity of each other were relaxed in the region starting from the Gly24 up to the Gly34 in order to allow for the additional motion amplitude for the PDZ2 flexible loop. The structure annealing algorithm was executed with 100'000 energy minimization steps for 1000 two-state conformers. A series of 1-9 state structure calculations were performed (Figure S3.2) indicating that a single state structure does not fulfill the experimental data well due to its high CYANA target function (TF), which is a measure of restraint violations, while 2 states appear to be sufficient to describe the experimental data (Figure S3.2). However, the relative population of the two states could not be identified with the use of the CYANA target function due to the low target function contrast for both apo and holo PDZ2 forms (Figure S3.5).

The twenty two-state conformers with the lowest TF were selected to represent the calculated two-state structure. They satisfy well the experimental restraints as indicated by the low CYANA target function (see Tables S3.1 and S3.2) and show well behaving Ramachandran plot statistics with less than 2% of the residues in the

disallowed regions (Table S3.3 and Figure S3.6). In addition, the resulting structures reproduce the known PDZ2 protein fold with root mean square deviation of 1.11 Å for apo and 1.34 Å for the complex from the corresponding reported crystal structures (Tables S3.1 and S3.2). Following a jack-knife procedure using the PDBcor software it was determined that the in the following presented correlations are experimentally overdetermined in the two-state apo structure as roughly 70% of the apo PDZ2 domain experimental distance restraints are required for the emergence of significant structural correlations (Figure S3.7).

The two-state structures of the PDZ2 domain of the apo and holo forms

The eNOE based two-state structures of the PDZ2 domain free (apo) and in complex with the peptide RA-GEF2 represented by twenty conformers for each state comprise overall the expected PDZ fold as expected (Figure 3.2a, 3.2b, 3.2d and 3.2e). When the two states are analyzed with the PDBcor software in standard settings [46] (Figure 3.2c and 3.2f) for apo and only in part also for the holo forms of the PDZ2 domain protein states are separatable for the β-sheet, α-helix 2, and the flexible loop Gly24-Gly34, and Chi1 angle values show for both apo and holo forms a characteristic variation between two state-dependent values reminiscent of local distinct configurations of the side chains (see Figures S3.3, S3.4). For both apo and holo PDZ2 forms the flexible loop exists in two conformations of which one is relatively loose and further apart from the binding site (dark blue state for the apo and red state for the holo form in Figure 3.2) and another one is relatively more confined and closer to the binding site (cyan state for the apo and yellow state for the holo form, please note for convenience the conformer separation and coloring will be kept consistent throughout the manuscript and we will refer to the state with the relatively more confined flexible loop that is closer to the binding site as state 1 and the other state as state 2).

The two separate states for the apo form are clearly separable from both the visual inspection (Figure 3.2a and 3.2b) as well as the correlation map of PDBcor (Figure 3.2c). Almost 60% of the entire protein domain appears to shuffle between two states with the most prominent structural difference around the peptide ligand binding site

comprising β-strand 2, α-helix 2 and the loop comprising residues Gly24-Gly34 but comprise the entire β-sheet to be discussed in details below. Structural correlation of the holo PDZ2 domain are much less pronounced as PDZ2 holo states are not correlated globally showing mainly minor local correlations (Figures 3.2f and S3.4). The only visible correlation of holo PDZ2 between the protein N-terminus and flexible loop comprising residues Gly24-Gly34 is probably spurious in nature as it is detected between two relatively flexible protein sites.

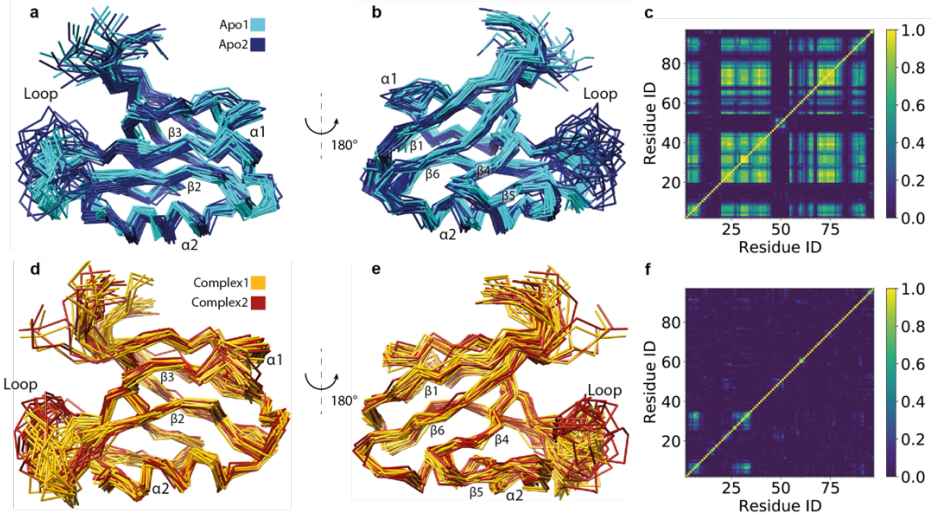

Figure 3.2: Two-state ensemble structures of the PDZ2 domain in the ligand-free apo form (a, b) and bound to the RA-GEF2 peptide (holo) (d, e) in two different orientations calculated with eNORA2 [33, 34] and CYANA software. Two apo states are colored in cyan and dark blue whereas two holo states are colored in yellow and red. The secondary structures and the loop comprising residues Gly24-Gly34 are indicated. Structural correlations for both protein ensembles were calculated with PDBcor in standard settings [46] and shown as distance correlation matrix heatmaps for the apo (c) and holo (f) forms of the PDZ2 domain.

Ligand induced conformational rearrangement of the PDZ2 domain

A systematic study of the conformational changes between the apo and holo PDZ2 scaffolds allows to gain insights into the binding mechanism. Conformational changes were first quantified in terms of the average distance between the apo and holo PDZ2 structures. For that both two-state structures were aligned to each other in UCSF Chimera [73], then C_α atom coordinates were extracted from all conformers and averaged to get a mean apo and mean holo structures where states A and B are averaged out. Next, a distance between C_α atom coordinates of the averaged apo and holo PDZ2 conformations was calculated for each residue. The residues that deviate more than 1.5 Å from each other were highlighted and mapped on the apo PDZ2 structure as shown in Figure 3.3. The loop comprising residues Gly24-Gly34 and the N-terminal flexible segment were excluded from the analysis due to their flexibility.

Conserved backbone conformational changes are concentrated to three sites. First site (S1) includes α-helix 1 and part of the β-strand 2 facing it. The second site (S2) includes the middle part of the β-sheet. The third site (S3) includes parts of the β-strand 5 and α-helix 2 in proximity of the flexible loop as summarized in Figure 3.3. The PDZ2 allosteric network spans from the RA-GEF2 binding site including residues Val75 and Val22 to the α-helix 1 over the S1 site, to the Lys54 over the S2 site and to the Val58 over the S3 site. Ligand binding yields a shift of the α-helix 1 and a part of the β-strand 2 facing it to allocate the ligand with a shift of the middle part of the β-sheet away from the binding site quantifiable also by the distance between the C_α atoms of residues Val22 and Val75 which is in the apo state 2 6.5 ± 0.3 Å, in the apo state 1 6.8 ± 0.4 Å versus 7.1 ± 0.5 Å in the PDZ2 complex. Aforementioned structural rearrangements are allosterically coupled to the ligand binding site, as backbone rearrangements between apo and complex PDZ2 domain are ligand induced a-priori. Those findings correlate with one of the major findings of Ranganathan et. al. who showed a statistical coupling between His71 and distal residues Ala46 and Ile52 that are part of the α-helix 1 and to the findings of Lee et. al. that indicated a coupling between residue Ile20 of the binding site and residues Ala39 and Val40 of the β-strand $\beta 3$ [1, 2]. Overall, a structural rather extensive and sophisticated correlation network of residues at atomic resolution in the coordinate space is identified that “feel” the ligand binding in line with previous indications [1, 2]. The mechanism is of induced fit-type as already previously reported by stopped-flow measurements [74].

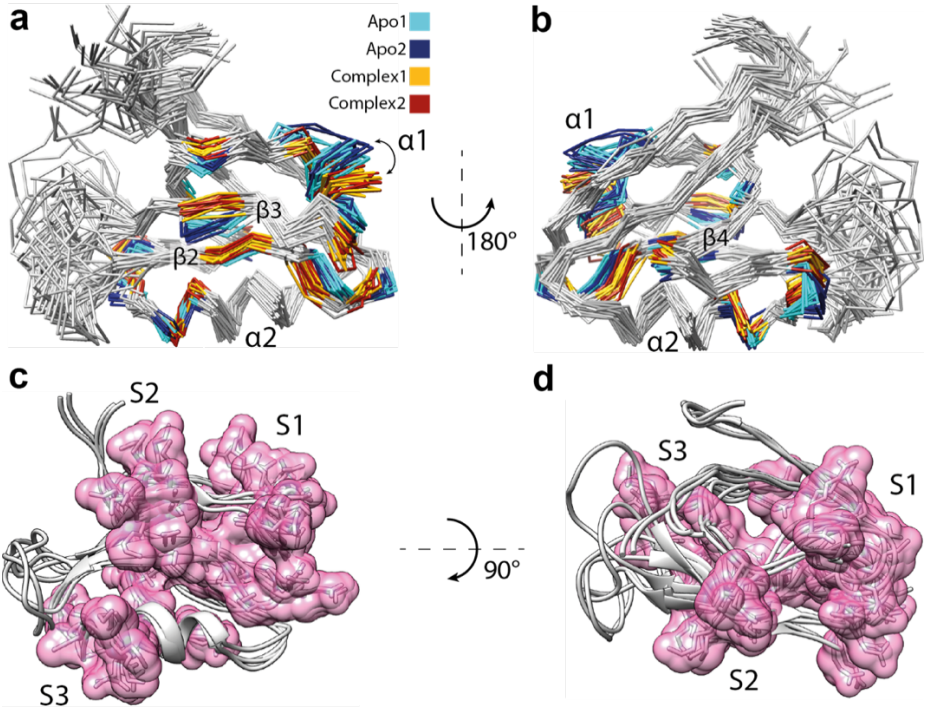

Figure 3.3: Ligand binding-induced structural changes of the PDZ2 domain. Top panel shows two views of the PDZ2 aligned apo-holo structural ensemble. All residues with apo-holo C_α-C_α distance of more than 1.5 Å except flexible regions are highlighted on the 3D protein structure with the color coding according to Figure 3.2 with cyan and blue representing the apo form and yellow and red the holo form, respectively (a, b). Ligand-induced allosteric movement of the α-helix 1 is indicated by the arrow. Significant rearrangements in the α-helix 1 and part of the β-strand 2 facing it, the middle part of the β-sheet and parts of the β-strand 5 and α-helix 1 in proximity of the flexible loop are validating previously reported allosteric interactions in the PDZ2 domain [1, 2]. Bottom panel shows three sites S1-S3 of the aforementioned allosteric network (c, d). The elucidated allosteric network spans from the binding site to the α-helix 1 over the S1, to the Lys54 over the S2 or to the Val58 over the S3.

Evidence for the conformational preselection in PDZ2 in terms of ligand binding

A further detailed investigation of both the apo and holo two state structures with focus on the sidechains of residues Ala69, Thr70, His71, and Lys38 close to the binding site suggests in part a conformational selection mechanism as both holo states are overlapping with apo state 1 (Figure 3.4, cyan). Apo state 2 features the sidechain of residues Ala69 and His71 pointing towards the flexible loop that is pushed further away

from the binding site and the sidechain of the residue Lys38 pointing directly to the binding site as shown in Figure 3.4a (blue). The role of Lys38 was further investigated by visualization of the PDZ2 molecular surface of the two representative states from apo and holo PDZ2 structures. Detailed analysis of the PDZ2 binding site conformation shows that the binding groove in the apo state 2 is obstructed by the sidechains of the residues Lys38 and Lys72 as shown in Figure 3.5. This finding suggests state 2 has to be a “closed” ligand-binding obstructing PDZ2 conformation while state 1 is the open ligand welcoming state that superimposes with the holo states indicating the presence of a conformational selection model for ligand binding.

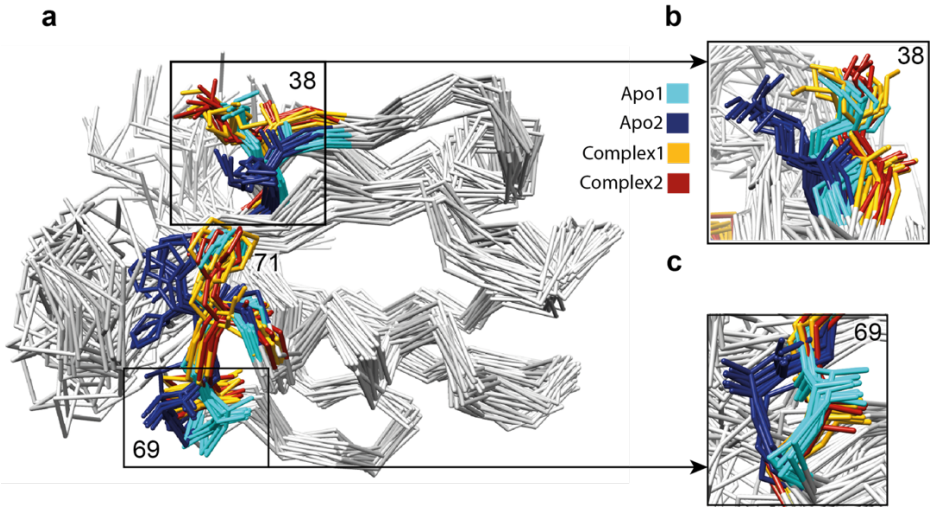

Figure 3.4: Conformational selection-based Ligand binding indicated by a structure comparison of the 2 state apo and holo structures. Sidechains of residues Ala69, Thr70, His71 (b) and Lys38 of both the 2 state apo and holo structures color coded as in Figure 3.2/3.3 (a) with inserts (b and c) show a super position of apo state 1 (cyan) with both the holo states (yellow and red) suggesting a conformational selection mechanism on PDZ2 for ligand binding.

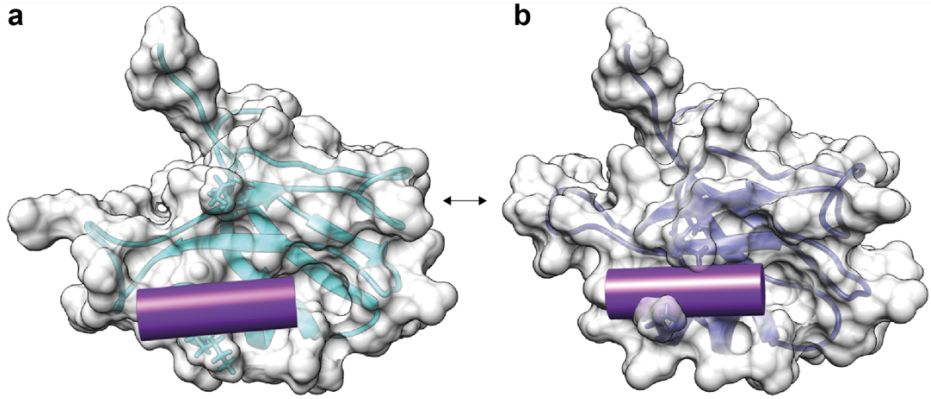

Figure 3.5: The apo form comprises an open ligand welcoming (a) and a closed ligand obstructing state (b)**.** The surface views of the two states of the PDZ2 of the apo form with state 1 (a) and state 2 (b) are shown with a ribbon representation and the important side chains of Lys38 and Lys72 shining through. The position of the RA-GEF2 peptide is visualized with a violet cylinder. It is visible that the access of the binding groove is obstructed in PDZ2 apo state 2 by sidechains of the residues Lys38 and Lys72 (b), which hints towards state preselection upon ligand binding giving raise to apo state 2 being the closed (b) and apo state 1 (a) the open conformation, respectively. This definition is in line with the super position of the apo states with the holo states shown in Figure 3.4 where apo state 1 is super imposing with the two states of the holo form.

An extensive correlation network within the apo PDZ2 domain steered by the dynamic loop

The above identification of an open ligand welcoming and a closed state of the apo PDZ2 is now analyzed within the entire protein domain by objective extraction of correlation with the PDBcor in standard settings [46]. The distance correlation matrix heatmap shown in the Figure 3.2c indicates that apo PDZ2 is a strongly correlated protein with correlations spanning throughout the protein fold with exception of the β-strand 1 and α-helix 1. The strongest correlations of the apo PDZ2 structural ensemble are concentrated to the RA-GEF2 binding site, β-strand 3 including residues Lys38, Lys72 and other residues involved in the conformational preselection including Ala69, Thr70 and His71 as shown in the Figure 3.6, but ca 60% of the entire protein domain is involved.

The analysis thereby indicates that the presence of the two states originates from the loop comprising residues Gly24-Gly34 which through its partial flexibility enabled by the two double glycine hinge motives Gly23-Gly24 and Gly33-Gly34 and its location at the edge of the protein structure comprising thermal intrinsic local dynamics,

moves such that in the apo state 1, the loop pushes sterically the side chain of Thr70 and His71 away enabling a shift of helix 2 closer to the loop. In addition, in its state 1 a steric push of the C-terminal end of the loop along with Ile35 induces a shift of the β-sheet via Val58/Leu59.

Within this context it is interesting to note that also in the case of the proline cis-trans isomerase human cyclophilin A, a loop having two hinge motives with each a double glycine motive is key for the two-state structure of the protein comprising more than 2/3 of the protein [37, 74] as in the case of the PDZ2 domain. While a generalization cannot be made from two cases only, the proposed mechanism of allostery, that is based on a dynamic loop feed by the thermic energy of the environment, which is sterically perturbing the folded part of the domain appears to be plausible. Because of its simplicity it thus also may be well presented in the protein world, which remains to be demonstrated.

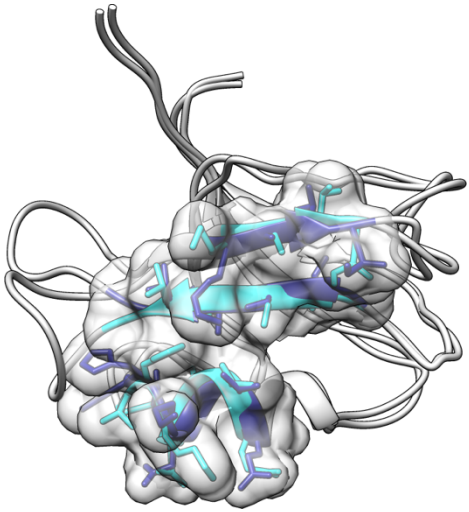

Figure 3.6: Core residues that are responsible for the two state correlations of the apo state are localized around the ligand binding site and the hydrophobic core of the protein. Fifteen highest correlated residues from the correlation matrix shown in the Figure 3.2c are highlighted as a single volume entity on the two-state apo PDZ2 structure including state-dependent ribbon coloring and side chain representation. As it is visible from the protein 3D structure highest correlations are concentrated to the protein binding site and all sites involved in the mentioned conformation preselection mechanism of Figures 3.4 and 3.5.

On the multi-level allosteric mechanism of the PDZ2 domain

The multi-state structures of the apo and holo form of the PDZ2 domain indicate at least two levels of protein allostery. The structural correlation of the apo form comprising an extensive structural correlation network between an open ligand welcoming and a closed ligand destructive state comprising roughly 60% of the entire domain are in line with the presence of a conformational selection allostery mechanism of ligand binding (Figure 3.6). In particular, the ligand binding site with α-helix 2 and β-strand 2, and the β-strand 3 are involved. Then, the ligand binding to the open state induces an extensive conformational change covering ~25% of the protein including again the binding site by definition as well as prominently α-helix 1 and the β-strand 4. Hence, with the induced fit step allosteric changes over the PDZ2 fold are spread. Interestingly the two allosteric networks are only partly overlapping. While both share the binding site, the apo form comprises an allosteric network with the entire β-sheet, while the holo form has with α-helix 1 and the β-strand 4 another allosteric network. Moreover, while both allosteric networks comprise the binding site they are structurally distinct also within the binding site. Finally, it is worth mentioning that the holo form comprises mainly one global structure with local plurality in the side chain configurations (i.e. distinct rotamers, Figure S3.4), which means that the ligand binding blocks long range structural correlations and plasticity (highlighted in Figure 3.7 by the same circle shape of the yellow and the red state).

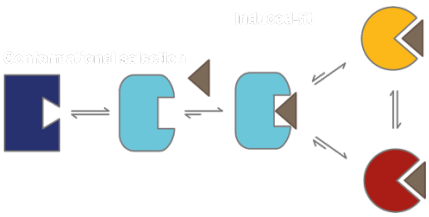

Figure 3.7: Cartoon on the multi-level allosteric mechanism of the PDZ2 domain, which can be summarized as a conformational selection between closed and open in the apo PDZ2 conformations followed by the induced fit mechanism upon binding to the one state that propagates allosteric rearrangements throughout the protein fold into the yellow/red state, which are distinct from each other mainly by local side chain rotamers and thus show the same shape in the cartoon. The color code of the PDZ2 domain as in the Figure 3.2 is followed and the ligand is indicated by a grey triangle.

Conclusions and Outlook

The presented work on the well-studied PDZ2 domain showcases the power of NMR with the high accuracy of the eNOEs that allows to solve multiple protein states at atomic resolution under physiological conditions in solution for the studies of protein allostery. It elucidated a two-level allosteric network at atomic resolution and pinpoints to the existence of structural rather extensive and sophisticated correlation networks of residues that in principle could be used for ligand binding regulation or signaling, that can “feel” the ligand binding, and that can be lost by ligand binding. In the context of the system of interest the PDZ2 allosteric binding mechanism was found to be combined from the broadly accepted induced-fit and conformational selection mechanisms. In more general terms, the presented work validates also in part previously reported allosteric indicators using a genetic algorithm [2] or experimental data [1]. It is furthermore obvious that such properties are to be expected in almost any biomolecular system awaiting to get explored.

Methods

Expression and purification of PDZ2 Domain

The DNA shuttle vector harboring PDZ2 sequence from human tyrosine phosphatase 1E (hPTP1E) was used for the bacterial expression of PDZ2 domain. The gene of interest included an N-terminal polyhistidine tag separated by an HRV-3C protease cleavage site. Expression and purification was carried out according to the previously reported procedures with minor modifications [1, 75]. Expression was performed in BL21 (DE3) *Escherichia coli* cells. The protein expression was induced after reaching OD_{600} of 0.8 with 1 mM IPTG. Stable isotope labeling was performed by resuspending cells in growth media supplemented with ^{15}N-enriched ammonium chloride and ^{13}C-enriched glucose. After overnight incubation with shaking, cells were harvested, resuspended and lysed with a Microfluidizer. The protein of interest was purified from the lysate with Ni-NTA chromatography. Then, the polyhistidine tag was cleaved with HRV-3C protease and the protein of interest was further separated by passing through the Ni-NTA column. The eluted protein was concentrated to 2 mM and the buffer was exchanged to a desired buffer for NMR (150 mM sodium chloride and 50 mM phosphate buffer at pH 6.8). A peptide (Ac-ENEQVSAV–COOH, BACHEM), or eight C-terminal residues from Rap Guanine Nucleotide Exchange Factor (RA-GEF2), was added to the final sample in concentration of 2 mM (1:1) for measurements of the PDZ2 domain bound to the ligand.

NMR experiments

The NMR measurements were performed on a 700 MHz Bruker spectrometer equipped with a triple resonance cryoprobe at 298 K. Processing and analysis of all NMR spectra was done with NMRPipe [76] and XEASY [77]. Structure calculations were done with eNORA2 within CYANA [34]. The rotational correlation times τ_c was calculated from the ^{15}N-relaxation experiments as described previously [60]. It was also optimized using a systematic screening approach with the goal of target function minimization in the structure calculations. Scalar couplings $^1J_{HNH\beta}$ were recorded as previously described [78] from a series of intensity-modulated HMQC spectra with 80($t_{1,max}$(^{15}N) = 28.2

ms)*512($t_{2,max}$(^{1}H) = 52.3 ms) complex points with an interscan delay of 1 s and 32 scans per increment. Scalar couplings $^1J_{H\alpha H\beta}$ were recorded as previously described [79] from a 3D ^{13}Ca-separated HACAHB-COSY experiment with 50($t_{1,max}$(^{13}C) = 14.2ms)* 54 ($t_{2,max}$(^{1}H) = 7.5 ms)* 2048 ($t_{3,max}$(^{1}H) = 204.9 ms) complex points, 16 scans per increment and 1s of interscan delay. Scalar couplings for aromatic side chain heavy atoms $^1J_{NC\gamma}$ and $^1J_{COC\gamma}$ were recorded as previously described [80] from intensity-modulated HSQC spectra with 200($t_{1,max}$(^{15}N) = 150.0 ms)*512($t_{2,max}$(^{1}H) = 51.2 ms) complex points, and interscan delay of 1.2 s and 16 scans per increment for $^1J_{NC\gamma}$ couplings and with 100($t_{1,max}$(^{15}N) = 75.0 ms)*512($t_{2,max}$(^{1}H) = 51.2 ms) complex points, and interscan delay of 1 s and 32 scans per increment for $^1J_{COC\gamma}$ couplings.

Structure calculation

The single and multistate protein structure calculation was done according to the previously reported procedure [35-37] using eNORA2 [33, 34] and CYANA [31]. Lower and upper distances from eNOEs, backbone, Hβ, and aromatic side-chain scalar couplings were used as inputs for the structure calculation. Calculations were done with 200'000 torsion angle dynamics steps for 100 conformers by simulated annealing. Identical heavy atoms from multistate conformers were kept together by a potential well with a bottom width of 1.2 Å as previously described [37].

eNOE dataset for PDZ2 domain in apo form

An exhaustive set of experimental restraints for the PDZ2 in the apo form consisted out of 1143 unidirectional distance restraints, 410 bidirectional distance restrains with highest precision of 0.1 Å and 65 scalar couplings that results in 17 restraints per residue as summarized in Table S3.1.

eNOE dataset for PDZ2 domain in holo form

For the PDZ2 in complex with RA-GEF2 peptide the experimental set of restraints consisted out of 995 unidirectional distance restraints, 489 bidirectional distance

restrains and 65 scalar couplings that results in 16 restraints per residue as summarized in Table S3.2.

Supplementary Information

Tables

NMR distance and dihedral constraints		
Distance constraints		
Total eNOEs	1553	
eNOEs from one pathway	1143	
eNOEs from two pathways	410	
Intra-residue, $\|i-j\|=0$	529	
Sequential, $\|i-j\|=1$	411	
Short-range, $\|i-j\|\leq 1$	940	
Medium-range, $3<\|i-j\|<5$	206	
Long-range, $\|i-j\|\geq 5$	407	
Dihedral angle restraints		
${}^{3}JHN_{\alpha}$ scalar coupling	65	
${}^{3}JH_{\alpha}H_{\beta}$ scalar coupling	55	
${}^{3}JHNCG$ scalar coupling (aromatic)	5	
${}^{3}JHNCOCG$ scalar coupling (aromatic)	5	
${}^{13}C_{\alpha}$ chemical shifts	86	
	One-state ensemble	Two-states ensemble

Structure statistics		
Average CYANA target function value (Å)	29.41 ± 0.05	7.81 ± 0.21
Violations		
Distance constraints (>0.5Å)	11	0
Dihedral angle constraints (>5°)	0	0
Deviations from idealized geometry		
RMSD (Å)		
Backbone to mean	0.42 ± 0.10	0.65 ± 0.09
Heavy atoms to mean	0.94 ± 0.11	1.27 ± 0.09
RMSD to X-ray structure (PDB ID 3LNX) / Å		
Backbone	1.32	1.11
Heavy atoms	1.79	1.86

Table S3.1 Structural statistics and CYANA input data for the apo PDZ2 domain.

NMR distance and dihedral constraints	
Distance constraints	
Total eNOEs	1484
eNOEs from one pathway	995
eNOEs from two pathways	489

Intra-residue, $\|i-j\|=0$	545	
Sequential, $\|i-j\|=1$	373	
Short-range, $\|i-j\|\leq 1$	918	
Medium-range, $1<\|i-j\|<5$	162	
Long-range, $\|i-j\|\geq 5$	404	
Dihedral angle restraints		
$^{3}JHN_{\alpha}$ scalar coupling	65	
$^{3}JH_{\alpha}H_{\beta}$ scalar coupling	55	
$^{3}JHNCG$ scalar coupling (aromatic)	5	
$^{3}JHNCOCG$ scalar coupling (aromatic)	5	
$^{13}C_{\alpha}$ chemical shifts	80	
	One-state ensemble	Two-states ensemble
Structure statistics		
Average CYANA target function value (Å)	25.87 ± 0.05	7.64 ± 0.19
Violations		
Distance constraints (>0.5Å)	7	1
Dihedral angle constraints (>5°)	0	0
Deviations from idealized geometry		

RMSD (Å)		
Backbone to mean	0.35 ± 0.07	0.65 ± 0.07
Heavy atoms to mean	0.87 ± 0.12	1.21 ± 0.08
RMSD to X-ray structure (PDB ID 3LNY) / Å		
Backbone	1.57	1.34
Heavy atoms	2.26	1.94

Table S3.2 Structural statistics and CYANA input data for the PDZ2 domain in complex with RA-GEF2 peptide.

	Apo	Complex
Most favored regions	64.9%	52.6%
Additionally allowed regions	31.5%	42.9%
Generously allowed regions	2.5%	3.3%
Disallowed regions	1.1%	1.2%

Table S3.3 Ramachandran statistics

Figures

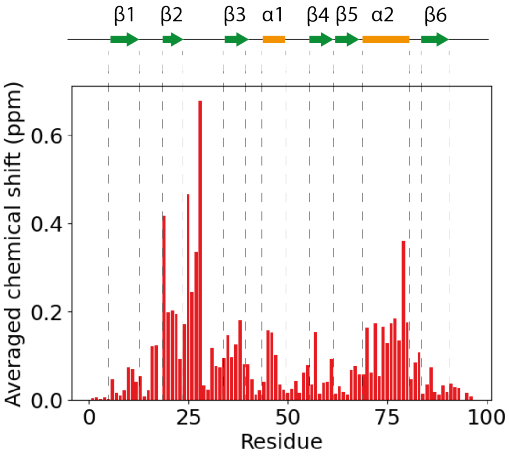

Figure S3.1: Ligand-binding induced chemical shift change *versus* the amino acid sequence of PDZ2 indicate the ligand binding site and allosteric sites. Ligand-induced ^{15}N and ^{1}H chemical shift changes (geometrically weighted with nitrogen shifts taken with weight of 0.3) measured in [^{15}N,1H]-HSQC spectra of free PDZ2 domain and in a 1:1 complex with the ligand peptide RA-FEF2.

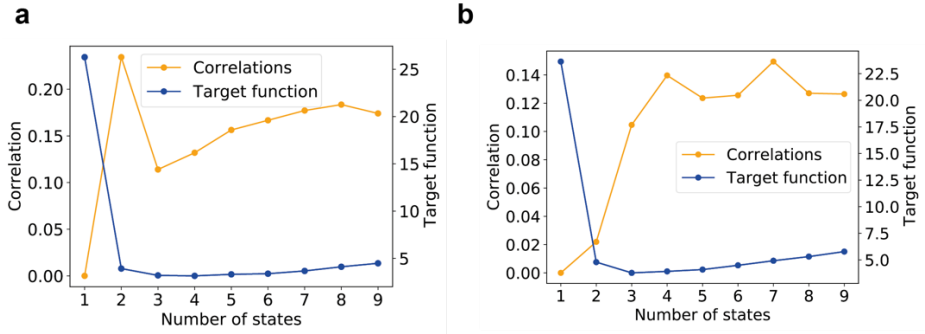

Figure S3.2: Two state validation of apo PDZ2 domain (a) and PDZ2 domain in complex with RA-GEF2 peptide (b). The CYANA target function (TF) values, which is the (weighted) sum of the squared violations of the conformational restraints versus number of simultaneously calculated states, is shown for 1-9 state structure calculations in blue. The importance of the ensemble-based structure determination is evident from the decrease of the TF with an increasing number of states and indicates that two states are sufficient to describe the experimental data well. In yellow the correlation value peak determined by the PDBcor calculator [46], which is also an indicator for the number of states of the system is shown. In the case of apo PDZ2 also 2 states are suggested as in the case of the CYANA TF, while in the case of the complex PDZ2 more states are predicted.

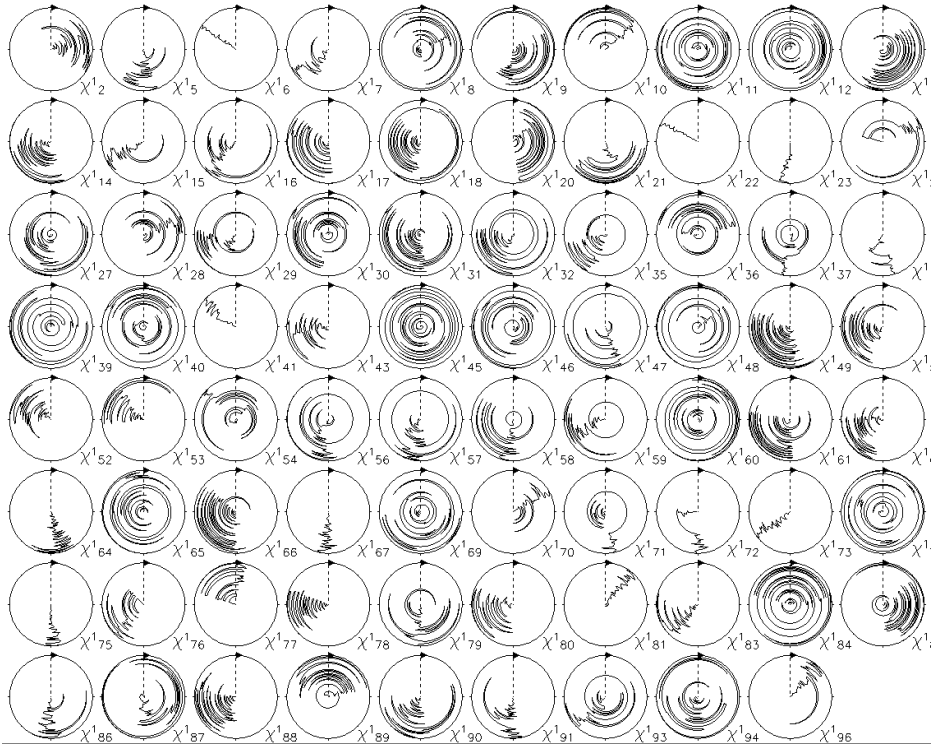

Figure S3.3: Chi1 angles of all the residues of the 40 conformers of the apo PDZ2 ensemble are shown in a circular plot.

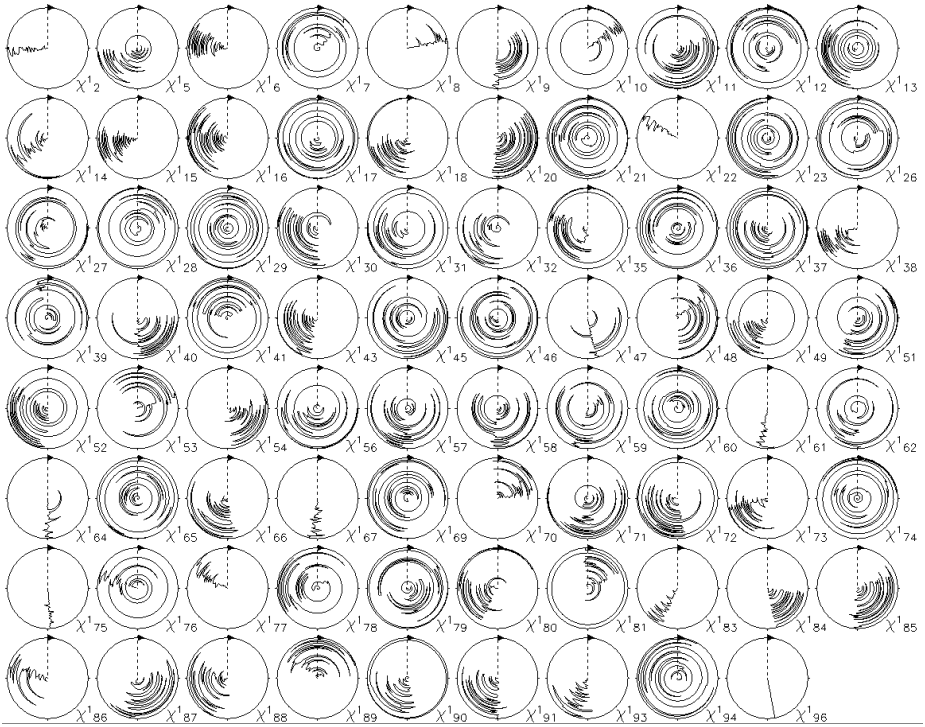

Figure S3.4: Chi1 angles of all the residues of the 40 conformers of the complex PDZ2 ensemble are shown in a circular plot.

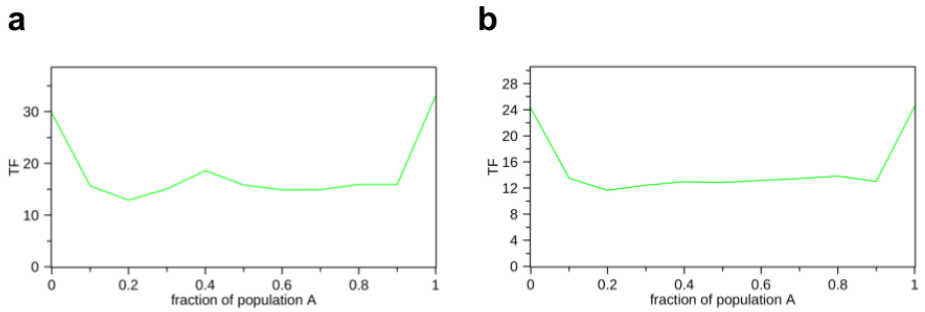

Figure S3.5: Population analysis of apo PDZ2 domain (a) and PDZ2 domain in complex with RA-GEF2 peptide (b) showing that the population could not be determined within the structure calculations. The graphs show the CYANA target function (TF) of the two-state structure calculations versus various populations. For this a pseudo ten-state structure calculation was set up allowing only two distinct states with various populations between 1:9 to 9:1 through symmetry restraints. From the Figure it is evident that the TF cannot determine the populations between 1:9 and 9:1.

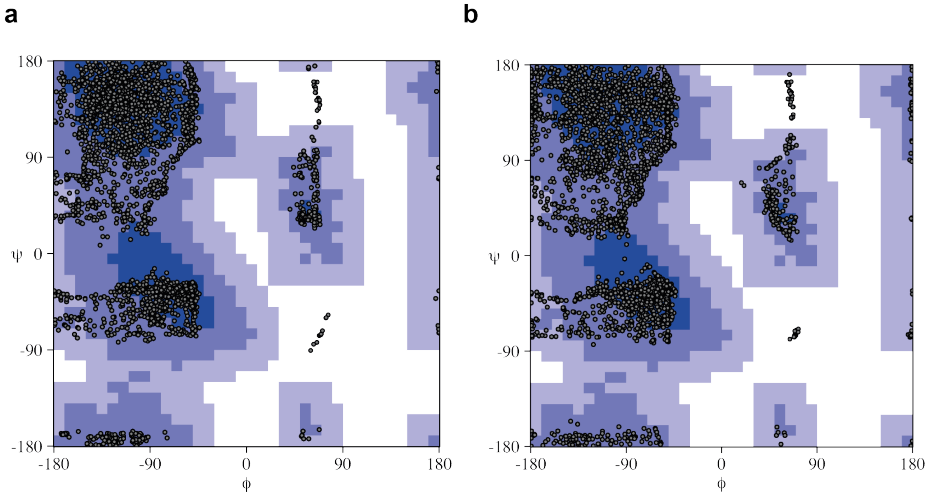

Figure S3.6: Ramachandran statistics of apo PDZ2 domain (all 2*20 conformers) (a) and PDZ2 domain in complex with RA-GEF2 peptide (all 2*20 conformers) (b).

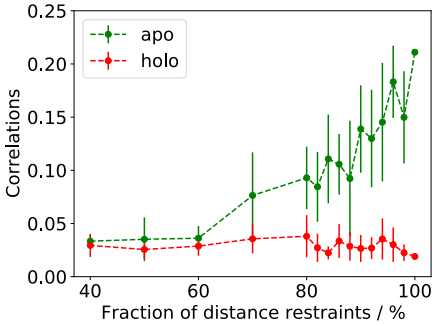

Figure S3.7: NOE network analysis and multi-state structure calculation stability studies. A series of two-state structure calculations of the PDZ2 domain in apo (green) and complex (red) form was performed with fractional distance restraint datasets and analyzed for structural correlations. Roughly 70% of the apo PDZ2 domain experimental distance restraints are required for the emergence of significant structural correlations between free PDZ2 states, whereas significant structural correlations in PDZ2 complex are not observed.

Chapter 4: Optimization and Validation of Multi-state NMR Protein Structures using Structural Correlations

This chapter is an adaptation from the manuscript under review in Journal of Biomolecular NMR: Dzmitry Ashkinadze, Harindranath Kadavath, Roland Riek*, Peter Güntert* Optimization and Validation of Multi-state NMR Protein Structures using Structural Correlations

Author's contribution: D.A. conducted experiments H.K. helped with scientific writing P.G. and R.R. supervised the project D.A., R.R. and P.G. wrote the manuscript. All authors discussed the results and contributed to the final manuscript

Introduction

Recently, we have developed the PDBcor software for the analysis of structural correlations in multi-state protein structures.[46] Within the PDBcor software, structural correlations indicating correlated motion are evaluated based on distance statistics in the protein bundle. These can be expressed as a matrix of correlation values between all residue pairs or alternatively as an overall correlation parameter (average correlation over the matrix). The correlation values calculated based on information theory represent the amount of information shared between protein residues in terms of their correspondence to the protein states. Correlations extracted with PDBcor are objective in the sense that they are not based on subjective structure superposition.

This work introduces structural correlations using PDBcor as a valuable quality control measure for multi-state protein structures in the context of NMR-based protein structure calculation.

An eNOE-based multi-state NMR structure calculation is based on the assumption that the protein of interest undergoes conformational exchange between different states. Under this assumption, the protein states are fitted to the data such that distances averaged over these states match best their experimentally measured values. The best fit corresponds to a minimal value of the target function, which is a weighted sum of the squared violations of the experimental distance restraints. In an initial step, a single-state NMR structure is calculated following standard procedure.[32] In single-state NMR structure calculations the target function along with bundle root-mean-square deviation (RMSD) and a list of violated distances in the calculated structure are good indicators to evaluate the accuracy of the protein structure and thus valuable tools for finding incorrect assignments/distance restraints.[26, 81] Next, multi-state structures are calculated that lead to the emergence of locally-split protein sites and networks. However, for multi-state structure calculations, the aforementioned tools lose their prominent role in finding erroneous restraints, because they may get “dissolved” in the additional degrees of freedom that come along with the multiple states. As we shall demonstrate, structural correlations provide an alternative additional quantitative parameter for multi-state structure validation that can be used for the final check of multi-state protein structures. As opposed to the target function that quantifies violations of

experimental restraints, structural correlations quantify the clustering separation between protein states and can be used to assess and quantify the features of the final multi-state protein structure. Therefore, structural correlations can be used to monitor and optimize the clustering separation between protein states that are not strictly dependent on the target function.

Selected assays together with corresponding demonstration examples were deposited at http://www.cyana.org/wiki/index.php/Tutorials

Results

Structural correlation value

All structural correlations in future sections are average correlation values that were extracted using PDBcor with default settings and the number of protein states set corresponding to the original CYANA protein calculation unless mentioned otherwise. The average correlation is the mean value of the elements of the distance correlation matrix A^d that is given as an output from the PDBcor software.

Optimization of the number of states

For a multi-state structure determination, the number of protein states that can be resolved meaningfully by the experimental restraints must be determined. The established procedure uses the target function decrease with the number of states calculated. The number of protein states is assessed by calculating protein ensembles with 1 to 9 states. The optimal number of states is then set according to the multi-state ensemble that achieves a minimum of the normalized target function or in other words to the minimum required number of states necessary to explain the experimental data. This is illustrated here for two previously reported model proteins, the WW domain of PIN1, yielding a two-state system, and GB3, yielding a four-state system, by multi-state calculations with previously reported procedures using CYANA and the published experimental restraints[36, 82] (Figure 4.1). We evaluated the ensembles with 1 to 9 states in terms of structural correlations using PDBcor. Structural correlations of the single-state ensembles were set to zero per definition as at least two states are required for the meaningful extraction of structural correlations. As it is clearly visible in Figure 4.1 the normalized target function reaches a minimum at the reported number of states in both cases and levels off with increasing number of states[36, 82]. As opposed to the target function, structural correlations of calculated structures do not plateau but show a maximum at the reported number of states. Arguably, structural correlations increase approaching the optimal number of states due to convergence of the protein bundle as required degrees of freedom become available and decrease afterwards as extra states start to fuse with existing states and make them statistically inseparable. Hence,

structural correlations provide an alternative method to determine the optimal number of states for multi-state structure calculation that can give more clear-cut results than the conventional target function-based analysis and are best used in concert with each other.

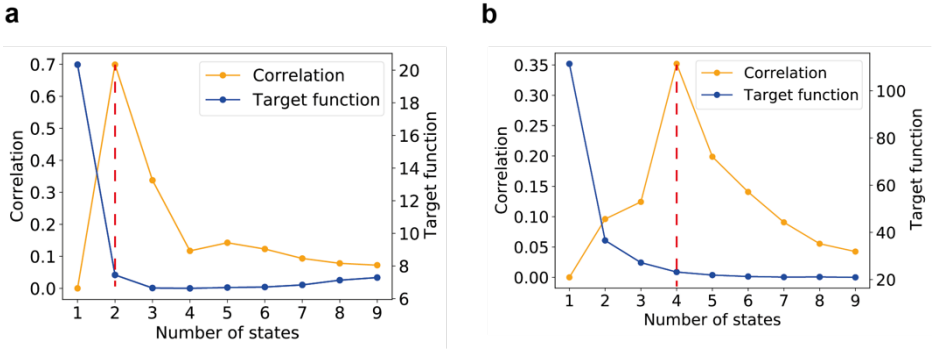

Figure 4.1 Results of the procedure to determine the optimal number of states for the WW domain (a) and the protein GB3 (b). The blue line represents the normalized target function and the orange line the average structural correlation as a function of the number of protein states. The correlation value peak at two states for the WW domain and four states for GB3, as determined previously on the basis of the normalized target function values. Nevertheless, structural correlations values are easier to interpret due to the prominent maximum at the optimal number of states.

Estimation of Protein State Populations

The large majority of documented multi-state protein structures feature a two-state model.[4] For such two-state models populations of individual states can be evaluated empirically by conducting a series of ten-state CYANA structure calculations in which the 10 individual states are separated in two controlled groups A and B.[36] Protein states in each group are tightly bound to each other. By varying the size of group A from 1 up to 9 conformers we can simulate a protein structure with population of state A rising from 10% up to 90%. In the established procedure optimal protein state populations are determined according to the minimum of the normalized target function.[82] Here, we present an estimation of protein state populations for two previously reported model proteins, the WW domain of PIN1 and cyclophilin A. In addition, for both systems protein ensembles calculated with varying population parameters were evaluated in terms of normalized target function values and structural correlations using PDBcor (Figure 4.2).

For the latter, two protein conformations representing both protein states were selected from each ten-state structure calculation and used as input to PDBcor making population analysis equivalent to the analysis of a series of two-state protein structures. Figure 4.2 shows that the target function approaches its minimum in range of 40–60% for both systems. Structural correlations for cyclophilin A exhibit a maximum at a state population of 50% with a slight shoulder at 20% and (equivalently) 80%. According to these observations the two protein states of cyclophilin A are populated equally or 20/80 judging by the correlation shoulder. Despite target function minimum at 40% structural correlations of the WW domain show a maximum at 10/90. However, it was also previously reported, that the estimation of protein state populations using the target function appears to be difficult for the WW domain.[36] While the correlation appears to be an alternative predictor of populations, it remains a difficult task.

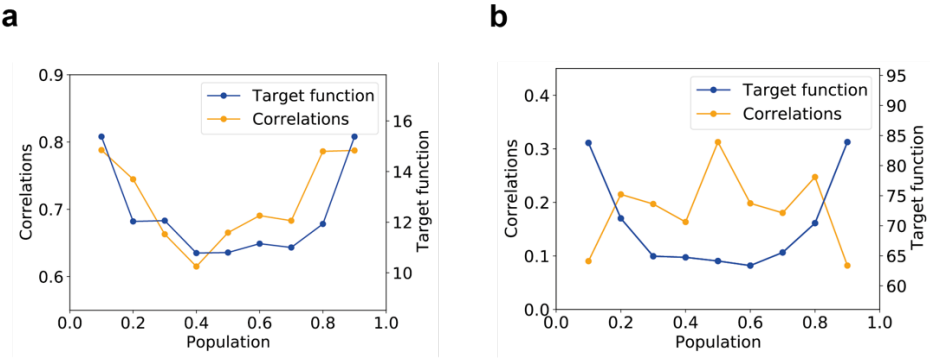

Figure 4.2 Results of population estimation studies for the WW domain (a) and protein cyclophilin A (b). The blue line corresponds to the normalized target function and the orange line to the average structural correlation as a function of the protein state A population.

Identification of Key Distance Restraints for Validation Purposes

In a multi-state structure determination, the identification of key eNOE distance restraints that reveal structural correlations is important in order to check their validity individually by inspection of the NMR spectra and analyses such as the NOE build-up rate quality. In order to find these key restraints individual distance restraints can be evaluated empirically in terms of structural correlations by calculating structures omitting a

particular distance restraint. As an example, a complete series of two-state structure calculations missing a particular long-range distance restraint was performed for the previously mentioned WW domain. Subsequently, each calculated bundle was evaluated for structural correlations and distance restraints were sorted in ascending order of the average structural correlation (Figure 4.3). A decrease in structural correlation caused by the removal of a particular distance restraint can either indicate that it is a key folding NOE restraint or a NOE restraint orchestrating correlated motion and protein states splitting. On the contrary, correlation increase due to removal of a particular eNOE could indicate potential structure calculation problems including distance restraint inaccuracy or misassignment. Distance restraints which removal contributed either to the twenty highest or twenty lowest correlation values were selected for further evaluation. In Figure 4.3b, key NOEs that were mapped onto the 3D structure are concentrated in the WW allosteric site and domain termini.[36] Nevertheless, since this approach evaluates contributions of individual NOEs, a possible contribution by distance restraints that are part of a redundant NOE subnetwork might be underestimated.

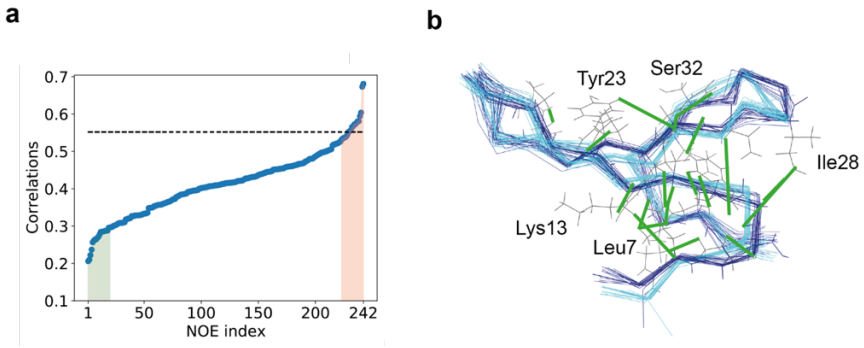

Figure 4.3 Results of the key distance restraint assay for the WW domain. (a) Long-range distance restraints from the WW domain were sorted according to the average structural correlation value obtained after their removal. The correlation level of the structure with all distance restraints is indicated by the black dashed line. Distance restraints corresponding to the twenty highest and twenty lowest correlation values are highlighted in red and green, respectively. (b) Twenty eNOEs invoking the biggest allosteric reduction (key eNOEs) are further illustrated on a two-state WW domain structure.

Validation of Individual Experimental Distance Restraints

In addition to the listing of key distance restraints that are important for the structural correlations, structural correlations obtained by the software PDBcor can also be used for the validation of individual distance restraints as illustrated for the two-state WW structure that was calculated once with and once without an upper limit distance restraint of 3.85 Å connecting the backbone amide H of Trp11 and HB2 of Asn26 (Figure 4.4). Structure bundles clearly indicate that the inclusion of this particular distance restraint affects locally the two-state separation of the side chain of Asn26. Nevertheless, the average target function value of the structure bundle including this distance restraint (6.52) does not favor it over the structure bundle lacking it (target function of 6.43). As opposed to the target function, the average structural correlation values clearly favor the calculation with this distance restraint (0.552) over the calculation without it (0.276).

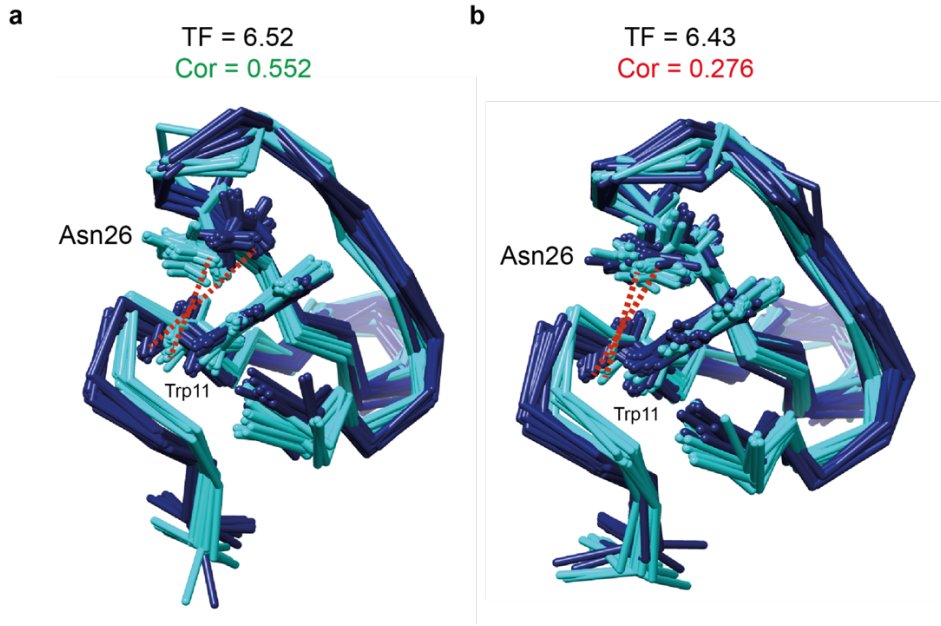

Figure 4.4 Selected region of the two-state WW domain structure bundles calculated with the previously reported protocol using CYANA and a set of experimental restraints acquired by liquid-state NMR.[36] Both structure bundles are colored according to the optimal two-state clustering deduced with PDBcor. Left structure bundle (a) was calculated using a full set of distance restraints, whereas a distance restraint of 3.85 Å connecting H of Trp11 and HB2 of Asn26 was excluded from the calculation of the structure bundle on the right (b). This distance restraint is depicted in both bundles as a red dotted line. Inclusion of the previously mentioned distance restraint clearly induces a local two-state separation for the involved residues. A positive effect of the inclusion of this distance restraint was detected by monitoring structural correlations, but not by the target function values.

Degree of Overdetermination of the NOE

In this section we discuss the collective effect of the NOE network as well as the degree of system overdetermination by monitoring structural correlations of protein bundles calculated from reduced distance restraint datasets. The three previously mentioned model systems (i.e. WW domain, protein GB3, and cyclophilin A) were evaluated for stability of the multi-state protein structure determination. All three models were calculated and analyzed for their reported number of 2, 4, and 2 states, respectively. A series of random subsets of the original experimental distance restraint dataset comprising from 20% to 90% of all available restraints was used as input for the structure calculations. Each experiment was repeated 10 times, always with new random fraction

of the dataset in order to counter the fact that not all distance restraints are equally important for the splitting of protein states. Calculated structure bundles were then evaluated for structural correlations. The correct fold was found for all three model proteins when supplied with 20% or more of the original distance restraints. Average structural correlations and their standard deviations for each dataset slice of the three model systems are shown in Figure 4.5. It is clearly visible that the WW domain has the most stable and most overdetermined NOE restraint system compared to other two proteins. Structural correlations or statistically significant splitting between protein states can be observed for any random fraction of the WW distance dataset that includes more than 50% of the original WW dataset, whereas for both cyclophilin A and GB3 more than 90% of original dataset are required for that.

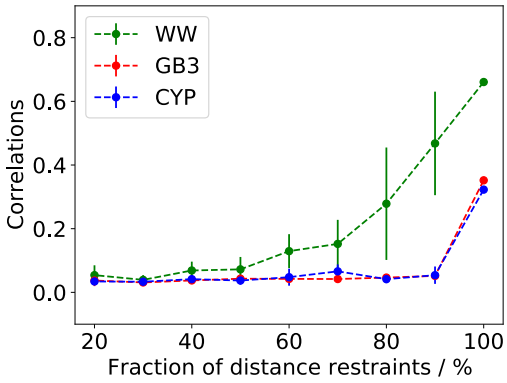

Figure 4.5 NOE network analysis and multi-state structure calculation stability studies. The two-state structure calculation of the WW domain is more stable than those of GB3 and cyclophilin A (CYP) according to the structural correlations of fractional distance restraint datasets since roughly 50% of the WW domain experimental distance restraints but over 90% of those for GB3 and cyclophilin A are required for comparable degrees of correlations.

Distance Range of Structural Correlations

We also studied how structural correlations derived from eNOE restraints depend on the distance between residues. Three deposited multi-state protein structure ensembles, including the WW domain of PIN1 (PDB ID 6SVC[36]), the protein GB3 (PDB ID 2LUM[82]) and cyclophilin A (PDB ID 2MZU[37]), were analyzed for structural correlations. For each system, all residue pairs were sorted in ascending order according to their average C^{α}–C^{α} distance in the published structures and separated into ten equal groups. Then, the average interresidual distance and the average correlation value was calculated for each group and plotted in Figure 4.6. Results show that the average

correlation values decrease with increasing distance between residues as it would be expected for local correlations that are limited in their span. Nevertheless, a certain level of structural correlations is retained throughout all distance groups as it would be expected for global correlations that are independent of the interresidual distance. Results also clearly indicate that correlated motion spans significantly larger distances than a single NOE (i.e. 5 Å), which can only be attributed to a collective influence of the NOE network.

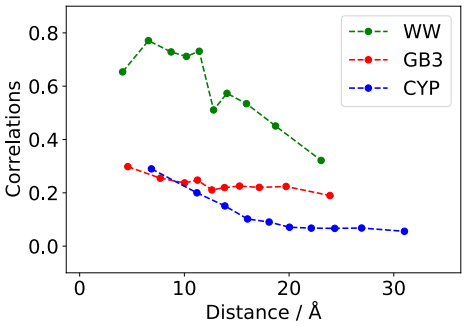

Figure 4.6 The distance dependence of structural correlations for the WW domain of PIN1 (green), the protein GB3 (red) and cyclophilin A (blue). Structural correlations of the WW domain and cyclophilin A experience a steeper decline as compared with protein GB3, which makes them more locally correlated than protein GB3. Significant correlation for distances above 5 Å (the maximum range of a single NOE) can only be explained by the effect of the NOE network.

Optimization of the CYANA Multi-State Structure

During the extensive testing of the above validation concepts, we also noticed that an insufficient number of conformers calculated with CYANA and an insufficient number of torsion angle dynamics steps can affect the structural correlation values through a suboptimal sampling by the calculated conformers. In particular, the number of torsion angle dynamics steps can have a major influence as shown in Figure 4.7. In order to illustrate the undersampling issue a series of two-state structure calculations of the WW domain were performed varying the number of torsion angle dynamics steps and the number of calculated conformers. Convergence was observed on the basis of structural correlations. The structure calculation convergence results summarized in Figure 4.7 indicated an adjustment of the CYANA calculation parameters to 200,000 torsion angle dynamics steps and 500 calculated conformers as optimal conditions for a multi-state structure determination.

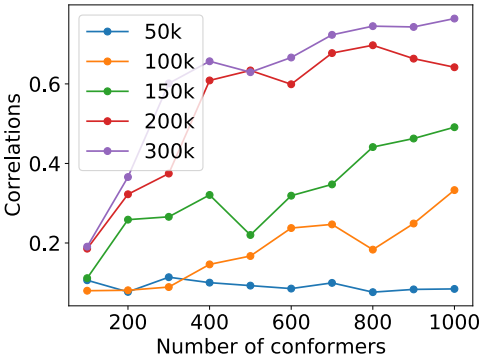

Figure 4.7 Screening of multi-state structure calculation conditions for the WW domain. Structural correlations of multi-state WW domain protein bundles indicate that conventional calculation of 100 conformers with 50000 torsion angle dynamics steps is not sufficient for convergence. Therefore, the basic structure calculation protocol was adjusted to 500 conformers with 200000 torsion angle dynamics steps.

Conclusions and Outlook

NMR-based multi-state structure determination is established[31, 35, 37] and has been demonstrated for 4 systems using eNOEs.[36-38, 82] The major remaining challenge that we identified in the protocol is the validation of the multi-state structures because the usual approach in standard structure calculations using the target function along with the list of remaining restrained violations[26, 32, 83] appeared not be sufficient to find all erroneous restraints or eNOE build-up curves, requesting detailed extensive manual analysis of individual restraints and NOE build-up fits along with many test calculations resulting in manually adapted, time-consuming and non-standardized procedures.

Here, we demonstrated that using the structural correlations obtained with the software PDBcor an additional tool for the validation of multi-state structure determinations that provide straight-forward information on the degree of overdetermination of the system is established, lists key restraints responsible for the identified structural correlations, and identified the number of states including their approximate populations necessary to fulfill the experimental restraints. Structural correlations are thus an important probe in the refinement stage of a multi-state structure calculation as they are sensitive to the protein state splitting, while the target function and the list of violated experimental restraints are important in earlier steps of the multi-state structure determination (in particular at the single-state and initial two-state structure determination phase). Together they constitute a powerful tool for the validation of NMR-based multi-state structures.

The PDBcor software for the calculation of structural correlations is freely available (https://github.com/dzmitryashkinadze/PDBcor).[46] PDBcor allows the straight-forward and objective determination of structural correlations in a given multi-state protein structure. The assays and subroutines performed and demonstrated here together with corresponding demonstration examples were deposited at http://www.cyana.org/wiki/index.php/Tutorials and can be straightforwardly adopted to individual systems. Together with the software package CYANA[32, 83] including the eNORA software[33, 34] for NOE build-up rate determinations, multi-state structures can be determined efficiently given NOESY cross peak assignments and intensities as an input. With these tools multi-state structures can be determined readily using eNOE

restraints. The additional NMR measurement time to acquire several (i.e. 3-4) combined $^{15}N,^{13}C$-resolved [$^{1}H,^{1}H$]-NOESY experiments instead of one is only approximately one week in order to obtain a multi-state structure that comprises the correlated dynamics of the protein of interest at atomic resolution and as such a unique quantitative information of presumably high biological relevance that currently no other technique than NMR can produce.

Methods

Protein structure calculations

eNOE-based multi-state structure calculations have been performed as reported previously for three proteins, the WW domain of PIN1 (PDB ID 6SVC[36]), the protein GB3 (PDB ID 2LUM[82]) and cyclophilin A (PDB ID 2MZU[37]). The experimental dataset for the WW domain[36] consists of 686 eNOE-derived distance restraints (271 bi-directional ones with 0% error and 415 uni-directional ones with 20% error) and 62 scalar couplings. The experimental dataset for the protein GB3[82] consists of 884 eNOE-derived distance restraints, 90 RDCs, and 201 scalar couplings. The experimental dataset for the protein cyclophilin A[37] consists of the 3640 eNOE-derived distance restraints, 396 RDCs, and 281 scalar couplings.

Structure calculations for this paper were executed following the established protocol[31, 35, 37] using eNORA2 for the spin diffusion correction[33, 34] and CYANA for structure annealing.[32, 83, 84] Upper and lower limit distance restraints produced by eNORA2, RDCs and scalar coupling restraints were used as input for multi-state structure calculations with CYANA. In each calculation 500 conformers were calculated with simulated annealing using 100,000 torsion angle dynamics steps per conformer. Corresponding heavy atoms from different states were kept together with the help of symmetry restraints in the form of a weak harmonic well potential with a bottom width of 1.2 Å.[31, 35] The twenty best conformers with the lowest final target function values were selected for structural correlation analysis.

Structural correlations

All structure correlations were extracted using the software PDBcor with default settings.[46] Each state of each conformer was provided as a separate protein entity as input for PDBcor. The number of states was set according to the CYANA calculation and the amplitude of thermal motion correction was set to 0.5 Å. Average correlation values were obtained as the mean value of the individual correlation values for each residue pair.

Chapter 5: Protein Allostery and Structural Correlations derived from Single-state NMR Structural Ensembles

This chapter is an adaptation from the manuscript in preparation: Dzmitry Ashkinadze, Piotr Klukowski, Harindranath Kadavath, Peter Güntert*, Roland Riek* Protein Allostery and Structural Correlations derived from Single-state NMR Structural Ensembles

Author's contribution: D.A. conducted experiments and wrote manuscript P.K. motivated to analyze PDB database H.K. helped with scientific writing P.G. and R.R. supervised the project. All authors discussed the results and contributed to the final manuscript

Introduction

NMR-based protein structures are relying on NOE distances that are of ensemble nature [19]. Even though an access to high quality average distances over the population of protein molecules in solution provides information that is sufficient to solve a multi-state protein structure [31], the majority of NMR protein structures deposited in the Protein Data Bank [7] are of a single-state nature. Since the amplitude of the protein motion is limited by the protein fold and NOE distance range, protein states acquired from the multi-state protein structure are typically partially overlapping. Considering also a limited precision of protein states it might be impossible to unambiguously discriminate between them. It was shown that in such cases a statistical study of the multi-state protein ensemble using software PDBcore can be used to efficiently extract information about so called structural correlations [46]. Protein regions are called correlated to each other if they move in a synchronized fashion or switch together between distinct local conformations in case if discrete system conformations are sampled.

The program PDBcor is able to elucidate structural correlations in an unbiased and automated way from the distance statistics of individual structural entities from the multi-state structure. It can uncover the protein regions that undergo synchronized motion, quantify correlations in structural ensembles and give insights into the biologically important correlated motion. PDBcor uses systematic clustering of protein conformers and information theory with aim to extract mutual information between individual residues [46].

Here we showcase that it is possible to extract meaningful structural correlations from the single-state protein NMR structures as the information about such interactions is not lost during conventional protein structure calculation despite the assumption of a single protein conformation. We also showcase the link between protein allostery and underlying protein motion by a systematic study of the correlated motion from the reported allosteric systems.

Results

Deviations from the mean structure in a single-state protein NMR ensemble are typically viewed in terms of experimental error, non-specific protein movement or protein bundle resolution. Nevertheless, we will show that prominent structural correlations between protein sites are conserved even if the protein structure was created as a single state that by design averages all input states in a single conformation. First, we will show the principle of correlation retention on an example of a minimal system (Figure 5.1). Then, on example of a synthetically created protein (Figure 5.2). Finally, we will show correlation retention on an example of the real experimental NMR-based structures (Figure 5.3).

Furthermore, since all deposited NMR structures might potentially incorporate information about correlated motion, we also investigated a protein data bank PDB and a database of allosteric proteins ASD to study the relation of correlated motion and allostery (Figure 5.4) [7, 9-12].

Principle of the Correlation Retention

Correlation retention can be shown on series of examples starting with a simple system consisting of interconnected atoms where each atom is equivalent and all bonds are of unit length such that this system exists in an equilibrium between two states as in Figure 5.1a. As shown in Figure 5.1a the system has a C2 symmetry along X-axis with an orange atom that can move horizontally and a varying angle α, which is a single degree of freedom of this hypothetical system and defines the configuration of the whole system by its value. In order to simulate the single-state protein structure calculation we need to assume that we have knowledge about the system architecture and that we have access to the average distances between all atom pairs. To simulate the final conformation, we need to calculate angle α that minimizes the target function or a sum of distance violations calculated as a sum of squared differences between an actual distance between atom pairs and its average value from two input conformations. Due to the simplicity of the input system, it can be easily solved numerically (Figure 5.1b). It turns out that by defining input angles as $\alpha_{\mathrm{in}} = \pm 0.39$ the target function approaches minimum

at values $\alpha_{out} \approx \pm 0.12$. This means that the final simulated conformations will be split between two solutions that are lying closer to each other than the input conformations, a situation that is typical for a real protein structure calculation.

So far, we described a single split structural element (Figure 5.1a). If we duplicate it as in Figure 5.1c and 5.1e we get a simplistic model with correlated orange atoms moving either in a synchronized (Figure 5.1c) or in an asynchronized (Figure 5.1e) mode. Both systems by design have two degrees of freedom as two angles that are coupled in the input data. Repeating the procedure that simulates protein structure solving we can numerically describe the target function as a function of two angles α_1 and α_2 (Figure 5.1d, 5.1f). It turns out that by defining input angles as $\alpha_{12} = \pm\frac{\pi}{8}$ the target function approaches minimum at all four conformations with correct two conformations having a deeper minimum peak of target function. A simulated single-state conformation calculation of this minimal model was able to find two correct conformations and therefore to keep the correlation between orange atoms in both synchronized and asynchronized models. Two false conformations corresponding to the other model and breaking the correlation were also detected with higher target function and therefore with population that is limited by the Boltzmann factor with energy gap proportional to the difference between target function minima. Results of this hypothetical experiment explain the mechanism behind partial retention of structural correlations in a single-state protein structure.

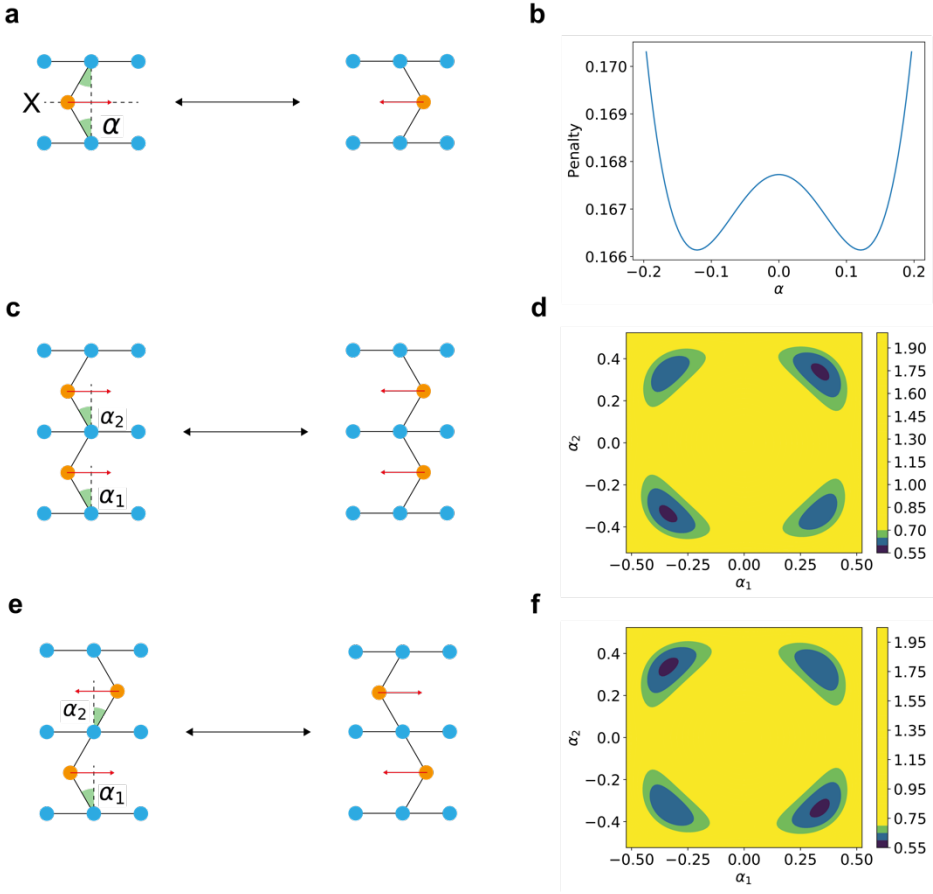

Figure 5.1 Theoretical study demonstrating the principle of correlation retention on an example of minimal systems. Each minimal system is presented in a dynamic equilibrium between two conformations (a, c, e). A set of average distances from each minimal model is used to evaluate possible conformations in terms of the target function. First model system (a) has only one degree of freedom (α) and a target function calculated for all possible values of α reveals two minima representing two optimized conformations of this system that are closer to each other than the initial conformations of the modeled system (b). Second and third model systems (c, e) have two degrees of freedom (α_1 and α_2) that are correlated to each other in the input data. Those two models are moving in a synchronized (c) or in an asynchronized (e) mode. Distance averaging and target function calculation along those two degrees of freedom (d, f) shows two minima representing two states of the input system and therefore allows to correctly reconstruct the input system conformations. In both cases two false states (target function difference between true and false states accounts for 0.013 in both models) that would break the correlation were also detected with higher target function and therefore with population that is limited by the Boltzmann factor with energy gap proportional to the difference between target function minima. This experiment illustrates the mechanism behind partial retention of the correlations in a single-state protein structure.

Validation of the Correlation Retention

In order to validate the retention of structural correlations a correlation model system on basis of protein cyclophilin A was constructed. For this a first PDB model was taken from the deposited liquid NMR bundle (PDB code: 2MZU) as a first state and a synthetical second state was created by manual tilting of two distant alpha-helices a1 and a2 and structure adjusting that removed a steric clash. Structure adjusting was made by calculating a synthetic set of distances in an adjusted cyclophilin A state and structure recalculation using CYANA and a standard structure calculation procedure [31]. The resulting synthetic dataset for protein cyclophilin A consisted of two protein states with correlated alpha-helices a1 and a2 (Figure 5.2a). Then, a synthetic peak list was created with CYANA covering all sidechain and backbone H^1-H^1 NOEs for distances in range between 0.1 Å and 5 Å such that peak intensities are corresponding to the state-averaged distances. There was not a single NOE directly connecting alpha-helices a1 and a2 in the generated peak list. Then, a conventional single state CYANA calculation was executed supplied with random selection of 500 NOESY sidechains peaks and 250 NOESY backbone peaks with 20% uncertainty for peak intensities to mimic a real structure calculation. The simulated structural ensemble is depicted in Figure 5.2b. As it can be seen from the structural ensemble the alpha-helices a1 and a2 are moving less compared to the initial structure as it would be expected from an averaged single-state structure. Conformers were optimally sorted in two states with PDBcor software and the separation between states in both alpha-helices is partially retained. Furthermore, structural correlations of the resulting structure bundle were also extracted with PDBcor in standard settings and number of states set to two. Structural correlations were summarized as a correlation heatmap over the protein sequence in the Figure 5.2c that confirms the correlation between alpha-helices a1 and a2 that was synthetically designed in the input data. Thus, the retention of long-range structural correlations in the single-state NMR protein structures was successfully validated.

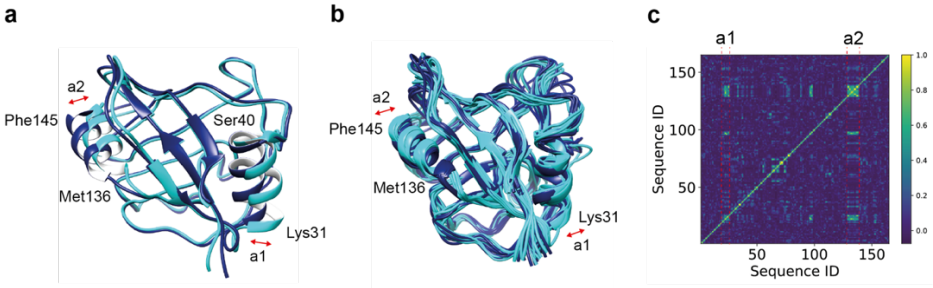

Figure 5.2 Validation of the correlation retention mechanism with a correlated data set that was synthetically created on the basis of the deposited cyclophilin A structure (a; PDB code: 2MZU). Two distant helices (distance larger than a single NOE) from cyclophilin were tilted and adjusted protein structure was recalculated to remove the electrostatic clash. Those two states are correlated due to the fact that the tilt on helix a1 is coupled the tilt of helix a2. Average distances were calculated from those two states and a randomly selected part of those distances was supplied with noise and used as input for a conventional single-state structure calculation in CYANA. The movement of alpha-helices a1 and a2 of the resulting structure (b) is much smaller compared to the initial structure as it would be expected from the averaged single-state structure, but the separation between states is partially retained. The final structure was further analyzed for structural correlations and initial synthetic correlation between helices a1 and a2 was detected (c). This validates that a single state structure calculation is able to retain information about the correlations between distant protein sites.

Single-dependent structural correlations of protein cyclophilin A

On the basis of reported correlated systems it is possible to confirm whether or not structural correlations extracted from a single-state structure are overlapping with structural correlations extracted from the multi-state structure. For this we calculated single-state and two-state structures of the model protein cyclophilin A and analyzed them for structural correlations. In order to make structural ensembles visually comparable 40 best single-state conformers together with 20 best two-state structures resulting in 40 individual conformers were aligned and depicted side by side for a two-state models of the cyclophilin. As it is visible from Figure 5.3, two-state ensemble of cyclophilin A has more deviation as it is expected from a multi-state structure due to the additional degrees of freedom. Conformers in presented ensembles were optimally sorted into states by PDBcor. Detailed examination shows that individual states have similar features in both single and multi-state ensembles. It is also visible that the states in multi-state simulation are equivalently populated as defined by the multi-state population, whereas in a single-state simulation states are not equivalently populated. Final examination of the correlation heat maps shows that correlations are highly similar

with correlations in the single state ensemble being less intense as correlations in the multi-state ensemble as it is also expected from a single-state structures.

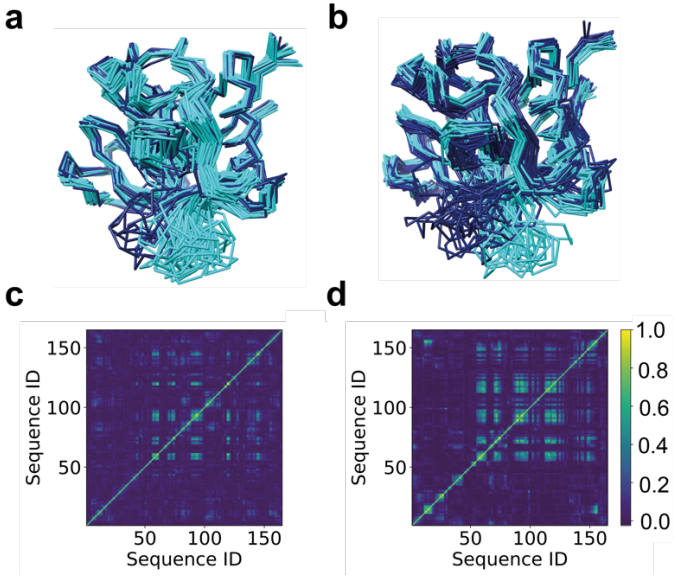

Figure 5.3 Comparison of the single-state structure coordinates (a), two-state structure coordinates (b), structural correlation heatmaps extracted from single-state structure (c) and structural correlation heatmaps extracted from two-state structure (d) for the protein cyclophilin A. It is visible that the two-state ensemble has more deviation than a single state ensemble. The states in two-state structure bundle are equivalently populated, whereas in single-state structure bundles states are not equivalently populated. Correlation heat maps are highly similar between a single and two-state structures with the single state structural correlations being less intense as multi-state structural correlations.

Exploration of the ASD Allosteric Database

The Allosteric Database was explored with aim to extract and analyze allosteric proteins with associated liquid NMR protein structures [7, 9-12]. The whole ASD database including 1949 allosteric proteins was cross-referenced with the UniProt database and 46 unique proteins with associated liquid NMR structures that are deposited at the PDB were identified [7, 85]. Majority of allosteric proteins are enzymes of big size that cannot be directly targeted with the liquid NMR due to fast relaxation [28]. Therefore, individual domains of such allosteric enzymes are typically studied with liquid NMR. The quality of the liquid NMR structures and limited allosteric site coverage further reduced our

selection down to only 12 structures that were finally selected to showcase the link between allostery and structural correlations.

An example of bovine calbindin D9K is summarized in Figure 5.4. Reported allosteric sites including Ca^{2+} binding site, N-terminal region and alpha-helix a3 are visualized in Figure 5.4a by the X-ray structure comparison of the metal-ion free structure (PDB code: 3ICB) and calbindin bound to Ca^{2+} and Mg^{2+} (PDB code: 1IG5) [86]. Those sites overlap with prominent structural correlations extracted from the liquid NMR structural ensemble (PDB code: 2MAZ) of apo bovine calbindin with PDBcor (Figure 5.4b). This example showcases that in some cases allosteric interactions that are based on protein correlated motion can be inferred by evaluation of the structural correlations directly from the single-state liquid NMR protein structures.

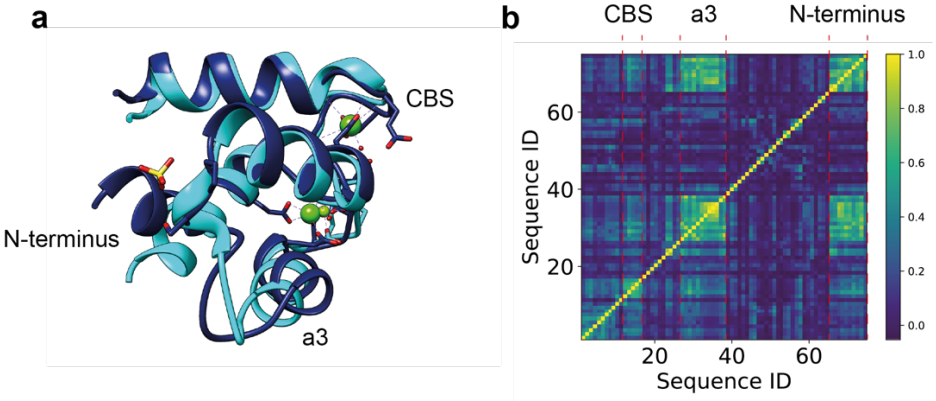

Figure 5.4 Overview of the protein bovine calbindin D9K in apo form (a; dark blue ribbon; PDB code: 3ICB) and in complex with Ca^{2+} and Mg^{2+} (a; cyan ribbon; PDB code: 1IG5) and the correlations map calculated from the liquid NMR structural ensemble (PDB code: 2MAZ) of apo bovine calbindin (b). Two reported allosteric sites located in the protein N-terminus and in the alpha α-helix 3 together with calcium binding site are visible both in the 3d structure of the protein ensemble and in the apo protein correlations heatmap [86].

Exploration of the PDB Databank

We analyzed 5077 liquid-state NMR structure ensembles from the protein data bank (PDB) that were represented as a bundle of at least 7 conformers [7]. All correlations were extracted with PDBcor using a two-state assumption. Extracted distance correlation

parameters show that a large population of deposited single-state NMR protein ensembles include structural correlations, as shown in Figure 5.5. This finding showcases that a majority of deposited structures encode an information about correlated motion that has to be systematically studied.

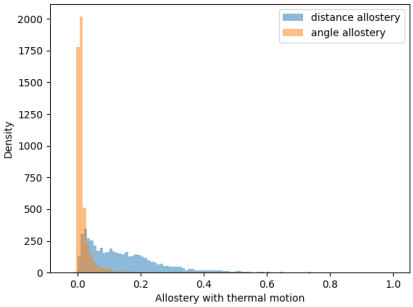

Figure 5.5 Extracted correlation parameters from the liquid-state NMR structure ensembles represented by a set of at least 7 conformers and deposited in the protein data bank.

Conclusions and Outlook

As it was shown on multiple examples of increasing complexity the single-state liquid NMR protein structures are able to describe proteins with conformational exchange. The limitation of the system degrees of freedom due to the faulty assumption of a single state makes protein states congested to each and reduces structure RMSD, but the information about the structural correlations between protein states is partially retained keeping prominent correlations encoded in the deviations of the single state structure. Further systematic investigation of the relation between protein correlated motion, structural correlations and protein allostery revealed that in some cases protein allostery can be inferred from structural correlations of the single state liquid NMR protein structures. The question of whether or not the analysis of a single domain of an allosteric enzyme acquired by liquid state NMR is relevant in the context of elucidating the global enzyme allostery remain unclear. In light of the potential link between protein allostery and structural correlations of the single-state liquid NMR protein structures we investigated structural correlations of all deposited liquid NMR structures from the protein data bank PDB and found that the majority of the deposited protein ensembles do have structural correlations.

Methods

Dataset for the protein Cyclophilin A

The multi-state eNOE-based structure calculation of the protein cyclophilin A have been reported previously (PDB ID 2MZU; [37]). The experimental dataset for the protein cyclophilin A consists of the 3640 eNOE-derived distance restraints, 281 scalar couplings and 396 RDC restraints.

Single-state structure calculation

Conventional single-state structure calculations were performed following the established protocol with software CYANA [31]. Upper limit distance restraints were produced with peak calibration procedure for a manually assigned peak list. In each calculation 500 conformers were calculated with simulated annealing using 100'000 torsion angle dynamics steps per conformer. Forty best conformers with the lowest final target function values were selected for further structural correlation analysis.

Exact NOE multi-state structure calculation

Exact NOE structure calculations were done according to the established protocol [31, 35, 37] using CYANA for structure annealing [32, 83, 84] and eNORA2 for the spin diffusion correction [33, 34]. Then a set of upper and lower limit distance restraints was produced by eNORA2. Distance restraints together with scalar coupling restraints were used as inputs for multi-state structure calculations. Each calculation resulted with 500 conformers that were calculated with simulated annealing using 100'000 torsion angle dynamics steps per conformer. Same heavy atoms from different state entities were kept together by symmetry restraints in the form of a weak harmonic well potential with a bottom width of 1.2 Å [31, 35, 37]. The best twenty conformers with the lowest final target function were selected for the further structural correlation analysis.

Structural correlations

Software PDBcor with default settings was used to extract all structural correlations (REF). Each protein entity was inputted to PDBcor as a separate PDB model. The number of states from the CYANA calculation was supplied to PDBcor and the thermal motion correction was executed at amplitude of 0.5 Å. Structural correlation values were obtained as the mean value from the distance correlation matrix outputted by PDBcor.

Conclusion and Outlook

Protein mechanisms of action have always been a highly challenging research topic. Despite the broadly available phenomenological observations of protein binding partners and various intra and intermolecular interactions there is a lack of mechanistic understanding behind protein mode of action. Increasing resolution of the methods in the field of protein biology allows us to resolve increasingly more features of the protein fold and protein dynamics that allow to deepen our understanding of protein-ligand interaction from the rigid body key-lock principle to more complex models based on protein thermodynamics.

Recent remarkable technological advancements of the cryo-EM allowed to bring down the resolution of the experimental protein structures and revolutionized the field of structural biology [22]. This led to the exponentially increasing pace of protein structure elucidations with cryo-EM. At the same time significant advances in the computational prediction of the protein structures with AlphaFold [23] allowed to make highly accurate prediction of protein structures. Aforementioned advancements in the field of protein structure determination drive the protein NMR research from the conventional structure elucidation towards protein dynamics field that allows to fully exploit unique advantages of NMR as it can investigate proteins, protein complexes and their intermediates in their native state with an abundance of the experimental techniques probing protein dynamics from which an eNOE technique is of a particular interest.

The exact NOE approach provides a unique way to solve multiple protein states at atomic resolution and therefore interpret the protein conformation space and correlated motion. The exact NOE approach is an improvement over the conventional protein structure elucidation with NMR due to the correction of the spin diffusion and extensive usage of the various NMR restraints including eNOE spin diffusion corrected distance restraints, RDC restraints, J-couplings and dihedral angle restrains. The adoption of the eNOE approach by the NMR community does not happen with a rapid pace presumably due to the long NMR acquisition time necessary to acquire multiple 3D-NOESY spectra and high required spectrum quality as is necessary to resolve large number of cross-peaks and generate large number of the distance restrains that would

overdetermine the NOE network and allow for the resolution of multiple states. Furthermore, demanding and highly specialized computational procedures are required to calculate multiple protein states. However, the Riek group actively investigates alternative spin diffusion correction algorithms based on a single NOESY spectrum and attempts to automatize multiple steps of the demanding multi-state NMR structure calculation with help of machine learning. For example, recent advances hint that in the near future it might be possible to solve an NMR protein structure with a single click and in a fully autonomous fashion by the application of the automated and spectrometer-integrated protein structure calculation software.

In my PhD work I studied the quantification of the correlated motion and protein allostery from existing highly accurate protein ensembles produced with exact NOE approach. The application of the machine learning allowed me to automatize the extraction of the valuable information about the correlated motion from the NMR protein ensembles and engineer the computational algorithm PDBcor that is more sensitive to the correlated motion compared to the conventional PCA-based algorithms.

Using this algorithm, it was possible to validate previously reported allosteric findings and obtain some novel structural insights from the exact NOE PDZ2 structures. Specifically, comparison of the apo and holo PDZ2 structures showed an allosteric interaction between the residue binding site and alpha helix 1, that was observed before with evolutionary method [2] and can be explained with induced-fit allosteric mechanism. Additionally, analysis of the PDZ2 apo states indicates that one of the states corresponds to the “open” form of the PDZ2 domain and one to the “closed” form in which the binding site is obstructed by the sidechains of residues Lys38 and Lys72. Moreover, according to the sidechains of the residues Lys38, His71 and Ala69 it was shown that the “closed” form is destabilized by the ligand binding. This observation shows that observed correlations between “open” and “closed” free PDZ2 states can be explained with conformational-selection allosteric mechanism.

Application of the PDBcor allowed to optimize some aspects of the liquid NMR multi-state protein structure elucidation as PDBcor provides an overall structural correlation value. If this value is zero it means that protein states are not correlated. It could indicate that either experimental restraints are not sufficient to separate protein

states or that there is absence of two states. High structural correlation value in turn means that it is possible to unambiguously distinguish between protein states. Therefore, it is possible to refer to the structural correlation value as a measure of separation between protein states that can be used as for a quantification of the multistate protein structures orthogonal to the CYANA target function. It was demonstrated that in some cases the CYANA target function is not sensitive towards certain key multi-state distance restraints, whereas the structural correlation value is. Furthermore, observation of the structural correlations enabled to find the optimal number of protein states easily due to the presence of the local maximum.

Moreover, increased sensitivity to the correlated motion of the PDBcor allowed us to see that correlated motion is a broad phenomenon even among single-state deposited protein ensembles. We proposed a theoretical model that explains the potential mechanism behind the retention of the correlations in the single-state structures and supply evidence of the single-state correlations for the previously reported allosteric protein structures. Further systematic investigation of the correlated motion gathered from the broad spectrum of deposited soluble protein structural ensembles might provide us with better understanding of the mechanism behind the protein motion and allow us to predict the protein motion for X-ray structures. At the time of writing the PDB databank contains 13451 liquid NMR structures from which there are 5076 protein structures represented by more than 7 conformers. If our assumption that the correlated motion can be extracted from the single-state structures is correct than those thousands of protein structures can be subjected to the broad exploratory analysis that might provide us with better understanding of protein correlated motion. First, it might be possible to map averaged correlated motion networks to the major evolutionary protein folds. Second, it might be possible to sort secondary structure elements and their combinations for their contribution to the correlated motion. Last, but not least thousands of the deposited structures and their correlation maps could be used as a dataset for the potential machine learning project targeted to the prediction of the correlated motion from the protein fold. If successful it could predict protein correlated motion from deposited protein X-ray structures.

Moreover, if the information about the protein correlated motion becomes broadly available it might be possible to connect it to the industrial setting by prediction of the protein active sites and potential drug candidates.

It was also shown that PDBcor is compatible with MD simulations. Further studies are required to understand whether or not PDBcor can improve the final protein trajectory analysis with its enhanced sensitivity over conventional PCA-based methods. However, since the current PDBcor implementation was created for the analysis of the NMR protein ensembles relatively high PDBcor computational cost might be a bottleneck for the analysis of the long MD trajectories and an alternative efficient MD-specific PDBcor alternative might be advantageous.

Appendix

Practical aspects of the exact NOE assignment

All details of the exact NOE assignment are described in the dedicated CYANA-Wiki tutorial.

Peak picking

Peak picking is performed as a first step of the exact NOE structure elucidation. Typically, NOESY spectrum with highest mixing time is picked as it has highest cross-peak intensities. Due to the practical issue of the overlapping peaks, eNOE peak picking is typically done with software NMRPipe [76].

According to the relaxation matrix analysis the cross-peak intensity in the first approximation depends linearly on the NOE mixing time and distance between hydrogens to the power of minus six [33].

$$\frac{\Delta M_{ij}(t)}{\Delta M_{ij}(0)} \sim \sigma_{ij} t \sim d_{ij}^{-6} t$$

Using this equation together with assumption that observed distances are in the NOE distance range as a prior knowledge it is possible to build an expectation model of a cross peak intensity as a function of the mixing time. Provided multiple NOESY spectra with different mixing times are measured it is possible to significantly purify the scope of peaked peaks by removing peaks with intensities that do not follow previously mentioned model as background peaks should not depend on the mixing time.

Literature

1. Fuentes, E.J., C.J. Der, and A.L. Lee, *Ligand-dependent dynamics and intramolecular signaling in a PDZ domain.* Journal of molecular biology, 2004. **335**(4): p. 1105-1115.
2. Lockless, S.W. and R. Ranganathan, *Evolutionarily conserved pathways of energetic connectivity in protein families.* Science, 1999. **286**(5438): p. 295-299.
3. Ishima, R. and D.A. Torchia, *Protein dynamics from NMR.* Nature structural biology, 2000. **7**(9): p. 740-743.
4. Bai, Y. and S.W. Englander, *Future directions in folding: The multi-state nature of protein structure.* Proteins: Structure, Function, and Bioinformatics, 1996. **24**(2): p. 145-151.
5. Karplus, M. and D.L. Weaver, *Protein-folding dynamics.* Nature, 1976. **260**(5550): p. 404-406.
6. Koshland Jr, D.E., *The key–lock theory and the induced fit theory.* Angewandte Chemie International Edition in English, 1995. **33**(23-24): p. 2375-2378.
7. Berman, H.M., et al., *The protein data bank.* Nucleic acids research, 2000. **28**(1): p. 235-242.
8. Swain, J.F. and L.M. Gierasch, *The changing landscape of protein allostery.* Current opinion in structural biology, 2006. **16**(1): p. 102-108.
9. Huang, Z., et al., *ASD v2.0: updated content and novel features focusing on allosteric regulation.* Nucleic Acids Research, 2013. **42**(D1): p. D510-D516.
10. Huang, Z., et al., *ASD: a comprehensive database of allosteric proteins and modulators.* Nucleic Acids Research, 2010. **39**(suppl_1): p. D663-D669.
11. Liu, X., et al., *Unraveling allosteric landscapes of allosterome with ASD.* Nucleic Acids Research, 2019. **48**(D1): p. D394-D401.
12. Shen, Q., et al., *ASD v3.0: unraveling allosteric regulation with structural mechanisms and biological networks.* Nucleic Acids Research, 2015. **44**(D1): p. D527-D535.
13. Green, S.M. and D. Shortle, *Patterns of nonadditivity between pairs of stability mutations in staphylococcal nuclease.* Biochemistry, 1993. **32**(38): p. 10131-10139.

14. Frauenfelder, H., et al., *The role of structure, energy landscape, dynamics, and allostery in the enzymatic function of myoglobin.* Proceedings of the National Academy of Sciences, 2001. **98**(5): p. 2370-2374.

15. Rod, T.H., J.L. Radkiewicz, and C.L. Brooks, *Correlated motion and the effect of distal mutations in dihydrofolate reductase.* Proceedings of the National Academy of Sciences, 2003. **100**(12): p. 6980-6985.

16. Bax, A. and S. Grzesiek, *Methodological advances in protein NMR.* Accounts of Chemical Research, 1993. **26**(4): p. 131-138.

17. Lipari, G. and A. Szabo, *Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity.* Journal of the American Chemical Society, 1982. **104**(17): p. 4546-4559.

18. Loria, J.P., M. Rance, and A.G. Palmer, *A relaxation-compensated Carr− Purcell− Meiboom− Gill sequence for characterizing chemical exchange by NMR spectroscopy.* Journal of the American Chemical Society, 1999. **121**(10): p. 2331-2332.

19. Wüthrich, K., *Protein structure determination in solution by NMR spectroscopy.* Journal of Biological Chemistry, 1990. **265**(36): p. 22059-22062.

20. Volkov, A.N., et al., *Solution structure and dynamics of the complex between cytochrome c and cytochrome c peroxidase determined by paramagnetic NMR.* Proceedings of the National Academy of Sciences, 2006. **103**(50): p. 18945-18950.

21. McConnell, H.M. and R.E. Robertson, *Isotropic nuclear resonance shifts.* The Journal of Chemical Physics, 1958. **29**(6): p. 1361-1365.

22. Li, X., et al., *Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM.* Nature methods, 2013. **10**(6): p. 584-590.

23. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.

24. Ortega, G., M. Pons, and O. Millet, *Protein functional dynamics in multiple timescales as studied by NMR spectroscopy.* Advances in protein chemistry and structural biology, 2013. **92**: p. 219-251.

25. Wüthrich, K., *NMR with proteins and nucleic acids.* Europhysics News, 1986. **17**(1): p. 11-13.

26. Güntert, P., *Automated NMR structure calculation with CYANA*, in *Protein NMR Techniques*. 2004, Springer. p. 353-378.

27. Clore, G.M., et al., *Impact of residual dipolar couplings on the accuracy of NMR structures determined from a minimal number of NOE restraints.* Journal of the American Chemical Society, 1999. **121**(27): p. 6513-6514.

28. Riek, R., et al., *Polarization transfer by cross-correlated relaxation in solution NMR with very large molecules.* Proceedings of the National Academy of Sciences, 1999. **96**(9): p. 4918-4923.

29. Iwahara, J., C.D. Schwieters, and G.M. Clore, *Ensemble approach for NMR structure refinement against 1H paramagnetic relaxation enhancement data arising from a flexible paramagnetic group attached to a macromolecule.* Journal of the American Chemical Society, 2004. **126**(18): p. 5879-5896.

30. Kumar, A. *Two-dimensional nuclear Overhauser effect in biomolecules*. in *Proceedings of the Indian Academy of Sciences-Chemical Sciences*. 1985. Springer.

31. Güntert, P., C. Mumenthaler, and K. Wüthrich, *Torsion angle dynamics for NMR structure calculation with the new program DYANA.* Journal of molecular biology, 1997. **273**(1): p. 283-298.

32. Güntert, P. and L. Buchner, *Combined automated NOE assignment and structure calculation with CYANA.* Journal of biomolecular NMR, 2015. **62**(4): p. 453-471.

33. Orts, J., B. Vögeli, and R. Riek, *Relaxation matrix analysis of spin diffusion for the NMR structure calculation with eNOEs.* Journal of chemical theory and computation, 2012. **8**(10): p. 3483-3492.

34. Strotz, D., et al., *ENORA2 exact NOE analysis program.* Journal of chemical theory and computation, 2017. **13**(9): p. 4336-4346.

35. Vögeli, B., et al., *Spatial elucidation of motion in proteins by ensemble-based structure calculation using exact NOEs.* Nature structural & molecular biology, 2012. **19**(10): p. 1053-1057.

36. Strotz, D., et al., *Protein allostery at atomic resolution.* Angewandte Chemie International Edition, 2020. **59**(49): p. 22132-22139.

37. Chi, C.N., et al., *Extending the eNOE data set of large proteins by evaluation of NOEs with unresolved diagonals.* Journal of biomolecular NMR, 2015. **62**(1): p. 63-69.

38. Vögeli, B., et al., *Exact distances and internal dynamics of perdeuterated ubiquitin from NOE buildups.* Journal of the American Chemical Society, 2009. **131**(47): p. 17215-17225.

39. Keepers, J.W. and T.L. James, *A theoretical study of distance determinations from NMR. Two-dimensional nuclear Overhauser effect spectra.* Journal of Magnetic Resonance (1969), 1984. **57**(3): p. 404-426.

40. Dalvit, C., et al., *WaterLOGSY as a method for primary NMR screening: practical aspects and range of applicability.* Journal of biomolecular NMR, 2001. **21**(4): p. 349-359.

41. Solomon, I., *Relaxation processes in a system of two spins.* Physical Review, 1955. **99**(2): p. 559.

42. Vögeli, B., *The nuclear Overhauser effect from a quantitative perspective.* Progress in nuclear magnetic resonance spectroscopy, 2014. **78**: p. 1-46.

43. Boelens, R., et al., *Iterative procedure for structure determination from proton-proton NOEs using a full relaxation matrix approach. Application to a DNA octamer.* Journal of Magnetic Resonance (1969), 1989. **82**(2): p. 290-308.

44. Linge, J.P., et al., *Correction of spin diffusion during iterative automated NOE assignment.* Journal of Magnetic Resonance, 2004. **167**(2): p. 334-342.

45. Schmidt, E. and P. Guntert, *A new algorithm for reliable and general NMR resonance assignment.* Journal of the American Chemical Society, 2012. **134**(30): p. 12817-12829.

46. Ashkinadze, D., et al., *PDBcor: An Automated Correlation Extraction Calculator for Multi-State Protein Structures.* Available at SSRN 3904349.

47. Palmer III, A.G., *NMR characterization of the dynamics of biomacromolecules.* Chemical reviews, 2004. **104**(8): p. 3623-3640.

48. Banerjee, S., et al., *2.3 Å resolution cryo-EM structure of human p97 and mechanism of allosteric inhibition.* Science, 2016. **351**(6275): p. 871-875.

49. Hekstra, D.R., et al., *Electric-field-stimulated protein mechanics.* Nature, 2016. **540**(7633): p. 400-405.

50. Bouvignies, G., et al., *Identification of slow correlated motions in proteins using residual dipolar and hydrogen-bond scalar couplings.* Proceedings of the National Academy of Sciences, 2005. **102**(39): p. 13885-13890.

51. Hummer, G., F. Schotte, and P.A. Anfinrud, *Unveiling functional protein motions with picosecond x-ray crystallography and molecular dynamics simulations.* Proceedings of the National Academy of Sciences, 2004. **101**(43): p. 15330-15334.

52. Nosé, S., *A molecular dynamics method for simulations in the canonical ensemble.* Molecular physics, 1984. **52**(2): p. 255-268.

53. McClendon, C.L., et al., *Quantifying correlations between allosteric sites in thermodynamic ensembles.* Journal of chemical theory and computation, 2009. **5**(9): p. 2486-2502.

54. Long, D. and R. Brüschweiler, *Atomistic kinetic model for population shift and allostery in biomolecules.* Journal of the American Chemical Society, 2011. **133**(46): p. 18999-19005.

55. Zhang, S., et al., *ProDy 2.0: Increased scale and scope after 10 years of protein dynamics modelling with Python.* Bioinformatics, 2021.

56. La Sala, G., et al., *Allosteric communication networks in proteins revealed through pocket crosstalk analysis.* ACS central science, 2017. **3**(9): p. 949-960.

57. Theobald, D.L. and D.S. Wuttke, *THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures.* Bioinformatics, 2006. **22**(17): p. 2171-2172.

58. Tiwari, S.P., et al., *WEBnm@ v2. 0: Web server and services for comparing protein flexibility.* BMC bioinformatics, 2014. **15**(1): p. 1-12.

59. Cock, P.J., et al., *Biopython: freely available Python tools for computational molecular biology and bioinformatics.* Bioinformatics, 2009. **25**(11): p. 1422-1423.

60. Kay, L.E., D.A. Torchia, and A. Bax, *Backbone dynamics of proteins as studied by nitrogen-15 inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease.* Biochemistry, 1989. **28**(23): p. 8972-8979.

61. Meyer, T., et al., *MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories.* Structure, 2010. **18**(11): p. 1399-1409.

62. Luque, F. and M. Orozco, *PCA suite: Software package for lossy trajectory compression using Principle Component Analysis techniques*. 2007, Molecular Recognition and Bioinformatics Group, University of Barcelona ….

63. McGibbon, R.T., et al., *MDTraj: a modern open library for the analysis of molecular dynamics trajectories.* Biophysical journal, 2015. **109**(8): p. 1528-1532.

64. Tonikian, R., et al., *A specificity map for the PDZ domain family.* PLoS biology, 2008. **6**(9): p. e239.

65. Walma, T., et al., *Structure, dynamics and binding characteristics of the second PDZ domain of PTP-BL.* Journal of molecular biology, 2002. **316**(5): p. 1101-1110.

66. Harris, B.Z. and W.A. Lim, *Mechanism and role of PDZ domains in signaling complex assembly.* Journal of cell science, 2001. **114**(18): p. 3219-3231.

67. Hung, A.Y. and M. Sheng, *PDZ domains: structural modules for protein complex assembly.* Journal of Biological Chemistry, 2002. **277**(8): p. 5699-5702.

68. Palmer, A., et al., *EphrinB phosphorylation and reverse signaling: regulation by Src kinases and PTP-BL phosphatase.* Molecular cell, 2002. **9**(4): p. 725-737.

69. Sato, T., et al., *FAP-1: a protein tyrosine phosphatase that associates with Fas.* Science, 1995. **268**(5209): p. 411-415.

70. Kozlov, G., et al., *Solution structure of the PDZ2 domain from cytosolic human phosphatase hPTP1E complexed with a peptide reveals contribution of the β2–β3 loop to PDZ domain–ligand interactions.* Journal of molecular biology, 2002. **320**(4): p. 813-820.

71. Kozlov, G., K. Gehring, and I. Ekiel, *Solution structure of the PDZ2 domain from human phosphatase hPTP1E and its interactions with C-terminal peptides from the Fas receptor.* Biochemistry, 2000. **39**(10): p. 2572-2580.

72. Doyle, D.A., et al., *Crystal structures of a complexed and peptide-free membrane protein–binding domain: molecular basis of peptide recognition by PDZ.* Cell, 1996. **85**(7): p. 1067-1076.

73. Pettersen, E.F., et al., *UCSF Chimera—a visualization system for exploratory research and analysis.* Journal of computational chemistry, 2004. **25**(13): p. 1605-1612.

74. Chi, C.N., et al., *A sequential binding mechanism in a PDZ domain.* Biochemistry, 2009. **48**(30): p. 7089-7097.

75. Bloem, R., et al., *Ligand binding studied by 2D IR spectroscopy using the azidohomoalanine label.* The Journal of Physical Chemistry B, 2012. **116**(46): p. 13705-13712.

76. Delaglio, F., et al., *NMRPipe: a multidimensional spectral processing system based on UNIX pipes.* Journal of biomolecular NMR, 1995. **6**(3): p. 277-293.

77. Bartels, C., et al., *The program XEASY for computer-supported NMR spectral analysis of biological macromolecules.* Journal of biomolecular NMR, 1995. **6**(1): p. 1-10.

78. Kuboniwa, H., et al., *Measurement of H N-H α J couplings in calcium-free calmodulin using new 2D and 3D water-flip-back methods.* Journal of biomolecular NMR, 1994. **4**(6): p. 871-878.

79. Grzesiek, S., et al., *Multiple-quantum line narrowing for measurement of H. alpha.-H. beta. J couplings in isotopically enriched proteins.* Journal of the American Chemical Society, 1995. **117**(19): p. 5312-5315.

80. Hu, J.-S., S. Grzesiek, and A. Bax, *Two-Dimensional NMR Methods for Determining χ1 Angles of Aromatic Residues in Proteins from Three-Bond JC 'Cγ and J NCγ Couplings.* Journal of the American Chemical Society, 1997. **119**(7): p. 1803-1804.

81. Buchner, L. and P. Güntert, *Systematic evaluation of combined automated NOE assignment and structure calculation with CYANA.* Journal of biomolecular NMR, 2015. **62**(1): p. 81-95.

82. Vögeli, B., et al., *The exact NOE as an alternative in ensemble structure determination.* Biophysical journal, 2016. **110**(1): p. 113-126.

83. Güntert, P., W. Braun, and K. Wüthrich, *Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA.* Journal of molecular biology, 1991. **217**(3): p. 517-530.

84. Herrmann, T., P. Güntert, and K. Wüthrich, *Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA.* Journal of molecular biology, 2002. **319**(1): p. 209-227.

85. Consortium, T.U., *UniProt: the universal protein knowledgebase in 2021.* Nucleic Acids Research, 2020. **49**(D1): p. D480-D489.

86. Andersson, M., et al., *Structural basis for the negative allostery between Ca2+-and Mg2+-binding in the intracellular Ca2+-receptor calbindin D9k.* Protein Science, 1997. **6**(6): p. 1139-1147.

Curriculum vitae

Personal Data

Name Dzmitry Ashkinadze

Birth Minsk, Belarus, May 15, 1994

Education

May 2018 – Dec. 2021	Doctoral Studies, ETHZ.
	Specialization: Bioinformatics, produced multiple papers
Sept. 2016 – Feb. 2018	Interdisciplinary Sciences MSc, ETHZ.
	Specialization: Biophysical chemistry; Good grades (5.1/6)
Sept. 2013 – Aug. 2016	Interdisciplinary Sciences BSc, ETHZ.
	Specialization: Physical Chemistry; Good grades (5.0/6)

Selected Publications

2021 Ashkinadze, Dzmitry, et al. “PDBcor: An Automated Correlation Extraction Calculator for Multi-State Protein Structures.“ Available at SSRN 3904349.

2021 Ashkinadze, Dzmitry, et al. “Optimization and validation of multi-state NMR protein structures using structural correlations.“ JB-NMR, in review.

Awards/Scholarships

Feb. 2019 – May 2019	Fellowship, Japan Society for the Promotion of Science (JSPS), Visiting fellow, Kyoto University, Kyoto, Japan.
Apr. 2015 – 2017	Scholarship, Anna und Franz Kägi Stifftung, Zurich, Switzerland.
Aug. 2016	Award, Theoretical Physics, Rudolf Ortvay Competition in Physics. Honorable Mention
Aug. 2011	Award, Physics, International Physics Olympiad, 2011, Bangkok, Thailand. Silver medal

Work Experience

May 2018 – Dec 2021 Research Assistant, ETHZ

- Solved multiple 3D protein structures and contributed to the software for the automated protein structure solving
- Developed and published a novel automated computational approach for the extraction of protein correlated motion from structural ensembles based on GMM clustering and information theory
- Actively collaborated with colleagues and produced a co-author publication
- Supervised an apprentice, led daily laboratory courses for 10-20 students and assisted in physical chemistry lectures
- Successfully completed a Project Management course

Jun. 2014 – May2018 Assistant of Physics Teacher, ExamPrep GMBH, part time, Zurich

- Assisted in physics lectures and prepared teaching materials
- Managed the online learning web platform

Skills

Program. Skills	Python, bash, R
Systems	Unix OS, Windows OS, Macintosh OS

Language skills

Russian	Mother tongue.
English	Fluent speaker, language level C1 verified by test TOEFL (95/120).
German	Fluent speaker, language level C1 verified by test DAF (17/20).

Interests / Hobbies

- Sports (sailing, hiking, skiing, chess and table tennis)
- Internet of things (IoT), built a radio