Università di Milano-Bicocca

Facoltà di Scienze Matematiche, Fisiche e Naturali

Dipartimento di Biotecnologie e Bioscienze

Dottorato di Ricerca in Biotecnologie Industriali – XXVII Ciclo

Network analysis and molecular dynamics simulations to investigate the link between structure and function in intrinsically disordered proteins and transcription factors

Tutor: Prof. Luca De Gioia

Cotutor: Dott. Elena Papaleo

Coordinator: Prof. Marco Vanoni

Matteo Lambrughi 075794

Anno accademico 2013-2014

Contents

Summary

Proteins are dynamic entities and their dynamics over different timescales are related to their structure and function. The understanding of the relationship between structure, dynamics and function is important not only for fundamental research in protein science, but it can have also impact on applications in biotechnological and pharmacological areas. In this thesis, we used a computational technique, Molecular Dynamics (MD), which allows to describe dynamics from the femto- to the millisecond timescale, to provide an atomic level description of target proteins related to human health. Recent developments in the MD field make now possible to study protein dynamics in details over a broad range of timescales, allowing to calculate properties that are in high agreement with the experimental observables. Nevertheless, conventional MD simulations are still far from being perfect and not always the physical model used to describe the system (i.e., the force field) is accurate enough for the case of study. Moreover, the protein risks to be entrapped in local minima for a long time in MD simulations without allowing to cover the whole conformational space. Thus, we integrated our simulations with data from different experimental biophysical techniques to overcome the limitations inherent in classical MD. Further, important conformational changes in the protein can be regulated by structural mechanisms propagated from a site distal to the site where the conformational change occurs, making them difficult to identify and describe. Our simulations were thus also analysed with methods inspired by graph theory, to unveil the effects induced over long distances in the protein. More in details, we selected, as model systems, protein domains that belong to the class of intrinsically disordered (ID) proteins (IDPs) and transcription factors. We both characterized the structural ensemble of these domains in solution and we investigated how the interaction with other biological partners or metal ions affects their structure and dynamics.

In Chapter 1, we introduced the basic concepts and theoretical basis of protein dynamics, the role that dynamics may play in cellular processes and protein function or even in structural alterations associated with disease. In this chapter, we also discussed the methods that we employed to analyse the simulations, with particular attention to principal component analysis and methods inspired by graph theory.

In Chapter 2, we presented the results of our work on ID domains, which are a class of protein that challenge the methods currently employed in structural biology. Indeed, IDPs are recently emerged as class of proteins that despite the lack of a folded thus IDPs do not conform to the classical structure-function paradigm. In particular, we applied the computational approaches detailed in Chapter 1 to provide a description of IDP heterogeneity in solution and to identify conformations that can transiently populate the free state of these ID domains and that resemble the conformations attained by the IDPs when they bind to their biological partners. More in details, Chapter 2 consists of four sub-chapters in which we discuss the data from our studies on different ID domains or in which we applied different methodologies.

In sub-Chapter 2.1, we focused on the kinase inhibitory domain (KID) of yeast Sic1, a cyclin-dependent kinase inhibitor (CKI) that regulates the progression of cell cycle in eukaryotic cells. Indeed, CKIs modulate the activity of cyclin-dependent kinases (Cdks) through the formation of ternary complexes with the kinase and its cognate cyclin. CKIs are IDPs that generally undergo folding upon binding to their partners. We characterized in atomic details the conformational ensemble of the free Sic1 KID fragment in solution by multi´-replicate MD simulations and spectroscopic and spectrometric data. Indeed, available experimental studies on Sic1 KID domain pointed out a propensity to populate both extended and compact conformations. We provide a description of this heterogeneous ensemble in solution, with particular attention to the compact states. We show that the KID domain is not merely disordered in the free state but part of the domain can attain helical structure, which correspond to the regions important for the interaction with the CdK-cyclin complex. We also identified a network of electrostatic interactions, which we speculated that could be involved in the stabilization of the more compact states of the domain.

In the subchapter 2.2 and subchapter 2.3, we discuss the results of our experimental and computational studies on the disordered region 182-291 (AT3$_{182\text{-}291}$) of human Ataxin-3, a deubiquitinating enzyme related to a neurodegenerative polyglutammine (PolyQ) disease. The AT3$_{182\text{-}291}$ flanks the polyQ tract and harbour two ubiquitin-interacting motifs (UIMs) that are involved in the interaction with ubiquitin. We integrated MD simulations with data collected by experimental biophysical spectroscopies and especially NMR spectroscopy thanks to a research period at the Structural Biology and NMR Laboratory at the Department of Biology of the University of Copenhagen (Copenhagen, Denmark). In subchapter 2.2 we provided a first model of the conformational states of the AT3$_{182\text{-}291}$ fragment in solution by multi-replicate MD simulations integrated to low resolution biophysical experimental data achieved by Circular Dichroism, Size-Exclusion Chromatography and Electrospray Ionization Mass Spectrometry (ESI-MS). This first characterization showed that AT3$_{182\text{-}291}$ is a monomeric ID domain in solution and is characterized by different states, ascribable to pre-molten globule populations with different degrees of compactness. We also identify a network of electrostatic interactions that might be involved in the regulation of tertiary propensity of the more compact states. We also showed that AT3$_{182\text{-}291}$ exerts protective effects against AT3 aggregation by ThT aggregation assays *in vitro*. We speculated that this effect might be induced by intermolecular interactions between AT3 and the UIMs of the AT3$_{182\text{-}291}$ fragment, in agreement with previous work that suggested a similar role of UIMs in the aggregation of polyQ-expanded huntingtin.

In subchapter 2.3, we achieved the first structural characterization at atomic resolution of AT3$_{182\text{-}291}$ in free state by NMR spectroscopy. We achieved an almost complete protein assignment of both backbone and side chain chemical shifts, confirming that AT3$_{182\text{-}291}$ is an ID domain, which is characterized by a high population of helical structures in four regions of the domain. Two of these tracts overlap with the two UIMs, showing that they can attain helical conformation even in absence of ubiquitin. We also investigated the possibility of transient long-range interactions in the AT3$_{182\text{-}291}$ fragment using site-directed spin-labelling and performing paramagnetic relaxation enhancement (PRE) experiments. We produced six different single point mutant variants of AT3$_{182\text{-}291}$ permitting a site-directed labelling able to cover all the protein sequence. The comparison of the PRE profiles pointed out that at least three different regions of AT3$_{182\text{-}291}$ are affected by the presence of the probe suggesting that these regions can make long-range contacts, whereas the C-terminal region is not involved in significant long-range contacts with the rest of the molecule. Interestingly, the regions involved in the long-range contacts comprise part of the two UIMs suggesting that also in the absence of ubiquitin the two UIMs can pair together, as observed for the AT3 UIMs in complex with Ubiquitin. Overall, the experimental data collected at atomic resolution by NMR and other data that were available on the system point out that the AT3$_{182\text{-}291}$ in solution is more extended and has very little propensity to populate collapsed states comparable to the ones that we observed by ESI-MS experiments and from the MD simulations collected in the first work. This observation highlights an important limitation of current MD force fields for the study of IDPs, which always encounter the risk of overcompation of the ID domain under investigation and thus makes even more important the integration of NMR and MD to study the IDP heterogeneous conformational landscape.

AT3 after the polyQ tract contains and additional disordered C-terminal part, which undergoes alternative splicing responsible for the main AT3 isoform, one of which is identified as 3UIM. Nothing is known from the structural point of view about the C-terminal region of AT3 3UIM isoform, but an additional UIM were predicted in its last 50 residues. It is not clear which is the role of the third UIM in the interaction with ubiquitin but it is involved in a multivalent binding to parkin ubiquitin ligase domain. In subchapter 2.4, we investigated the C-terminal region of AT3 3UIM isoform employing Replica-Exchange MD (REMD) simulations and two different force fields and two different solvent models. The results obtained with the two force fields are not in agreement and show relevant differences in the helical propensity and in the compactness of AT3 3UIM. We compared our results with NMR data available, concluding that the simulations performed with CHARMM22* force field and the TIPS3P water model give a reasonable description of the helical content for AT3 3UIM, but unrealistic overcollapsed structures. These results provide one more example of the limitation of current MD force fields in describing IDPs conformational states, a problem which can be mitigated by the development of *ad-hoc* force fields for IDPs. Despite these limitations, we provided a first description of the conformational states of the third UIM of AT3 showing that it has a low propensity to populate helical conformations in solution. We suggest that the location of suboptimal residues for the geometry of a α helix in the 3UIM can determine its low propensity to populate stable helical conformations in solution. We also identified, in 3UIM free ensemble, conformations that resemble the NMR structure of a canonical UIM in complex with ubiquitin. Due to the low propensity for helical states in solution, further studies are necessary to clarify if 3UIM is non canonical UIM and thus bind with very low affinity or cannot even bind ubiquitin.

In Chapter 3, we presented the other main subject of this thesis, i.e. the role of native metal cofactors or the effects induced by non-essential metals on structure and dynamics of protein domains from transcription cofactors, such as the zinc finger domain of Superman protein and the DNA binding domain of the tumor suppressor p53. P53 plays a central role in cellular functions, through the recruitment of multiple partners, but there is a very limited knowledge of the effects induced on p53 structure and function upon binding of other biomolecules, like the DNA. In Subchapter 3.1, we presented the MD study that we carried out on p53 with the scope to identify regions distal from the DNA binding site that are allosterically modulated by DNA interaction and can modulate the capabilities of p53 to recruit biological cofactors. We employed both canonical and enhanced-sampling atomistic simulations and analysed these by graph theory-inspired methods to study the p53 DNA binding domain (DBD) in the absence and presence of DNA. We described conformational changes and long-range intramolecular communication paths in p53 DBD upon DNA interaction. We showed that DNA binding promotes a conformational change in the loop S6-S7, a region 3 nm away from the DNA binding site of p53, increasing of more than four folds a population of a minor state, which is also present in the free protein in solution. We identified an interface of p53 that binds biological partners related to p53 transcription-independent functions that is not accessible in the states induced by DNA interactions. Moreover, we identified 7 binding partner for p53 DBD for which the interactions is likely to occur in the proximity of S6-S7 loop and can be regulated by the allosteric changes induced by DNA. An example of these interactors is Ku70, involved in apoptotic signalling. Our data suggest a mechanism to protect p53 from interactions with partners that elicit its transcription-independent apoptotic signalling when p53 is bound to the DNA and has to carry out its transcription functions.

In the Subchapter 3.2, we focused on the study of the effects induced by cadmium, a known pollutant metal and carcinogen, on zinc-binding proteins. The structural characterizations of zinc-binding proteins and of the alterations associated with cadmium toxicity are essential to understand the role of metals in these biological relevant problems. In this context since the lacking of extensive experimental analysis, computational approaches can be highly valuable in structural investigation of zinc-binding protein but approximations are present in the force field for the description of metal coordination and in the treatment of cadmium. We developed a protocol based on classical molecular mechanics (MM) and quantum chemical (QM) calculations to derive optimized force field parameters for the metal ions, which can then be used in MD simulations. The protocol has been designed to describe the coordination with zinc or cadmium of p53 and other metal binding proteins. At first, we provided a description of the protocol and the steps used to optimize force-field parameters using simulating annealing to fit the MM calculations to the QM minimum energy geometry, and in addition to fit the dissociation binding energy curves. We applied our protocol to obtain optimized parameters for the coordination site of zinc finger domain that is composed by two cysteines and two histidines (Zn-Cys_sHis_s). We tested the behaviour of our optimized parameters in MD simulations of Superman protein zinc finger domain and compared with MD simulations performed with the classical CHARMM22/CMAP force field. We pointed out the effectiveness of our approach and optimized parameters that give results in good agreement with experimental structure and QM calculations. We also developed new parameters for cadmium in Cys_sHis_2 coordination geometry. We showed that our first set results better than the parameters present in literature, showing conformational alterations similar at those experimentally observed, but a further optimization are ongoing to accurate describe the molecular mechanisms of cadmium toxicity.

The thesis closes in Chapter 5, in which the conclusions about the project are discussed pointing out that MD simulations performed with state-of-the-art force fields, if integrated with experimental biophysical techniques, are a powerful and effective tool to study protein dynamics and their relationship with structure and functions.

Riassunto

Le proteine sono entità dinamiche e sono caratterizzate da moti su diverse scale temporali strettamente legati alla loro struttura e funzione. La comprensione della relazione tra struttura, dinamica e funzione è quindi importante non solo per la ricerca di base nel campo dello studio delle proteine, ma può avere anche un impatto applicativo in ambito biotecnologico e farmacologico. In questa tesi, abbiamo utilizzato una tecnica computazionale, la Dinamica Molecolare (MD), che consente di descrivere dinamiche su scale di tempi che spaziano dal femto ai millisecondi, fornendo una descrizione a livello atomico di proteine importanti per i loro risvolti sulla salute umana. Recenti sviluppi nel campo MD rendono ora possibile studiare in dettaglio la dinamica delle proteine, permettendo di calcolare proprietà in forte accordo con le osservabili sperimentali e di avere accesso a informazioni su una vasta scala di tempi. Tuttavia, simulazioni MD convenzionali sono ancora lontane dall'essere una tecnica perfetta e non sempre il modello fisico utilizzato per descrivere il sistema (campo di forza) è sufficientemente accurato per il caso in studio, nonché durante tali simulazioni il sistema può rimanere intrappolato in minimi energetici locali, impedendo di campionare in maniera accurata lo spazio conformazionale accessibile al sistema. Nei lavori presentati in questa tesi di dottorato abbiamo quindi integrato le nostre simulazioni MD con i dati provenienti da diverse tecniche sperimentali biofisiche per superare i limiti intrinseci nella MD classica. Inoltre, importanti fluttuazioni conformazionali delle proteine possono essere regolate da meccanismi strutturali che si propagano da una regione lontana dal sito in cui si verifica il cambiamento conformazionale, il che le rende difficili da identificare e da descrivere. Abbiamo quindi utilizzato metodi ispirati alla teoria dei grafi per analizzare le nostre simulazioni, in maniera da descrivere effetti indotti a lungo raggio nelle proteine. Più in dettaglio, abbiamo selezionato, come sistemi modello, i domini di proteine che appartengono alla classe delle proteine intrinsecamente disordinate (IDPs) e di fattori di trascrizione. Abbiamo caratterizzato sia l'*ensemble* strutturale di questi singoli domini in soluzione, sia abbiamo studiato come l'interazione con altri partner biologici o ioni metallici incida sulla loro struttura e dinamica.

Nel capitolo 1, abbiamo introdotto i concetti di base e le basi teoriche della dinamica delle proteine, il ruolo che possono svolgere le dinamiche nei processi cellulari e nella funzione delle proteine o anche in alterazioni strutturali associate a patologie. In questo capitolo, abbiamo anche discusso i metodi che abbiamo impiegato per analizzare le simulazioni, con particolare attenzione all'analisi delle componenti principali del moto e ai metodi ispirati alla teoria dei grafi.

Nel capitolo 2 sono presentati i risultati del nostro lavoro condotto sui domini intrinsecamente disordinati (ID), che sono tra i casi più difficili di studio nel campo proteine. Le IDP sono una classe di numerose proteine che sono state recentemente identificate e che sebbene non possiedano una struttura tridimensionale precisa e ben organizzata quando si trovano libere in soluzione, svolgono comunque

funzioni biologiche fondamentali nelle cellule, non seguendo quindi il classico paradigma struttura-funzione. Quindi le differenze sostanziali che intercorrono tra le IDPs e le proteine con una struttura tridimensionale definita rendono difficile studiarle con i metodi attuali di biologia strutturale. In particolare nel nostro lavoro abbiamo applicato gli approcci computazionali illustrati nel capitolo 1 per fornire una descrizione dell'eterogeneità conformazionale delle IDP in soluzione e per individuare conformazioni che possono essere transitoriamente popolate da tali domini nel loro stato libero e che ricordano le conformazioni che essi assumono quando si legano ai loro partner biologici. Più in dettaglio, il capitolo 2 si compone di quattro sotto-capitoli in cui sono discussi i dati ottenuti dai nostri studi su diversi domini ID o in cui abbiamo applicato metodologie diverse.

Nel sotto-capitolo 2.1, ci siamo concentrati sul dominio inibitorio delle chinasi (KID) della proteina di lievito Sic1, un inibitore delle chinasi ciclina-dipendenti (CKI) che regola la progressione del ciclo cellulare nelle cellule eucariotiche. Infatti, le CKI modulano l'attività delle chinasi ciclina-dipendenti (CDK) attraverso la formazione di complessi ternari con la chinasi e la ciclina associata. Le CKI sono IDPs che vanno incontro a processi di *folding upon binding* a seguito dell'interazione con i loro target. Abbiamo caratterizzato a livello atomico l'*ensemble* conformazionale del frammento KID di Sic1 in soluzione attraverso simulazioni MD e dati provenienti da tecniche spettroscopiche e spettrometriche. Infatti, precedenti studi sperimentali effettuati sul dominio KID di Sic1 mostrano una propensione ad assumere conformazioni sia estese che compatte. Abbiamo fornito una descrizione di questo *ensemble* eterogeneo in soluzione, con particolare attenzione alla descrizione degli stati compatti. Abbiamo dimostrato che il dominio KID non è semplicemente disordinato nel suo stato libero, ma parte del dominio può assumere strutture a elica, che corrispondono alle regioni importanti per l'interazione con il complesso CDK-ciclina. Abbiamo anche individuato un network di interazioni elettrostatiche, che abbiamo ipotizzato potrebbero essere coinvolto nella stabilizzazione degli stati più compatti del dominio.

Nel sotto-capitolo 2.2 e 2.3, discutiamo i risultati dei nostri studi sperimentali e computazionali sulla regione disordinata 182-291 ($AT3_{182-291}$) dell'Atassina 3 umana, un enzima deubiquitinante correlato a una malattia neurodegenerativa da espansione di poliglutammine (PolyQ). Il dominio $AT3_{182-291}$ fiancheggia il tratto polyQ e contiene due *Ubiquitin Interacting Motifs* (UIMS) che sono coinvolti nell'interazione con l'ubiquitina. Abbiamo integrato simulazioni MD con dati sperimentali raccolti mediante tecniche spettroscopiche e specialmente via spettroscopia NMR, grazie ad un periodo di ricerca presso lo Structural Biology and NMR Laboratory nel Dipartimento di Biologia dell'Università di Copenaghen (Copenhagen, Danimarca). Nel sottocapitolo 2.2, abbiamo fornito un primo modello degli stati conformazionali del frammento $AT3_{182-291}$ in soluzione mediante più di un microsecondo di simulazioni MD integrate a bassa risoluzione con dati sperimentali ottenuti da tecniche biofisiche quali dicroismo circolare, cromatografia ad esclusione e Spettrometria di Massa (ESI-MS). Questa prima caratterizzazione ha dimostrato che AT3$_{182\text{-}291}$ è un dominio ID monomerico in soluzione ed è caratterizzato da diversi stati strutturali, ascrivibili a popolazioni *pre-molten globule* con diversi gradi di compattezza. Abbiamo inoltre identificato un network di interazioni elettrostatiche che potrebbero essere coinvolte nella regolazione della propensione ad assumere struttura terziaria degli stati più compatti. Mediante saggi di aggregazione della ThT abbiamo anche dimostrato che AT3$_{182\text{-}291}$ esercita effetti protettivi contro l'aggregazione dell'Atassina 3. Abbiamo ipotizzato che questo effetto potrebbe essere indotto da interazioni intermolecolari tra l'AT3 e i motivi UIM del dominio AT3$_{182\text{-}291}$ in accordo con un precedente lavoro che ha suggerito un ruolo simile delle UIMs nell'aggregazione dell'huntingtina, un'altra proteina poliQ.

Nel sottocapitolo 2.3, abbiamo ottenuto la prima caratterizzazione strutturale a risoluzione atomica di AT3$_{182\text{-}291}$ in stato libero mediante spettroscopia NMR. Abbiamo ottenuto un'assegnazione quasi completa dei *chemical shifts* degli atomi della catena principale e delle catene laterali dei residui di tale dominio, confermando che AT3$_{182\text{-}291}$ è un dominio ID caratterizzato da una elevata frazione di strutture elicali in quattro regioni del dominio. Due di questi tratti si sovrappongono con le due UIM, dimostrando che essi possono assumere conformazione ad elica, anche in assenza di ubiquitina. Abbiamo anche studiato la possibilità della presenza di interazioni a lungo raggio transienti tra differenti regioni del dominio AT3$_{182\text{-}291}$ mediante tecniche di *spin-labelling* e l'esecuzione di esperimenti NMR (*paramagnetic relaxation enhanchement experiments*, PRE). Abbiamo prodotto sei differenti varianti mutante di AT3$_{182\text{-}291}$ per consentire una marcatura sito specifica che fosse in grado di coprire tutta la sequenza della proteina. Il confronto dei profili PRE ha mostrato che almeno tre diverse regioni del dominio AT3$_{182\text{-}291}$ sono influenzate dalla presenza della sonda, suggerendo che queste regioni possono stabilire contatti a lungo raggio, mentre la regione C-terminale non è coinvolta in interazioni significative con il resto della molecola. Le regioni coinvolte in contatti a lungo raggio comprendono parte delle due UIM, suggerendo che anche in assenza di ubiquitina le due UIM si possono appaiare insieme, come è stato osservato per le UIM in complesso con ubiquitina. Globalmente i dati sperimentali raccolti mediante spettroscopia NMR e altri dati che erano già disponibili sul sistema sottolineano che il dominio AT3$_{182\text{-}291}$ in soluzione è esteso e ha molto poca propensione a popolare stati collassati paragonabili a quelli che abbiamo osservato da esperimenti ESI-MS e dalle simulazioni MD raccolte nel primo lavoro. Questa osservazione evidenzia un'importante limitazione di campi di forza MD attuali per lo studio delle IDP, che incontrano il rischio di overcompattare il dominio ID in esame, rendendo quindi ancora più importante e necessaria l'integrazione della spettroscopia NMR e simulazioni MD per studiare accuratamente l'eterogeneo *ensemble* conformazionale delle IDP.

L'AT3 dopo il tratto polyQ contiene un'ulteriore regione C-terminale disordinata che subisce *splicing* alternativo, responsabile della produzione delle principali isoforme dell'AT3, uno delle quali è identificata come 3UIM. Nulla è noto dal punto di vista strutturale sulla regione C-terminale dell'isoforma AT3 3UIM, ma è stato proposto che un'ulteriore UIM sia localizzata nei suoi ultimi 50 residui. Non è chiaro quale sia il ruolo della terza UIM nell'interazione con ubiquitina, ma è noto che sia coinvolto in un nel legame polivalente con un dominio della parkina. Nel sottocapitolo 2.4 abbiamo studiato la regione C-terminale dell'isoforma AT3 3UIM impiegando simulazioni *Replica Exchange* (REMD) e usando due campi di forza diversi e due modelli di solvente differenti. Infatti la caratterizzazione dell'*ensemble* di strutture delle diverse varianti di una proteina è fondamentale per una migliore comprensione della proprietà funzionali e non funzionali della proteina stessa. I risultati ottenuti con i due campi di forza sono in disaccordo e mostrano differenze rilevanti nella propensione ad elica e nella compattezza della regione AT3 3UIM. Abbiamo confrontato i nostri risultati con i dati NMR disponibili in letteratura, concludendo che le simulazioni effettuate con il campo di forza CHARMM22* e il modello di solvente TIPS3P danno una descrizione ragionevole del contenuto in elica per AT3 3UIM, ma mostrano anche strutture overcompatte e irrealistiche. Questi risultati forniscono un ulteriore esempio delle limitazioni dei campi di forza MD attuali nella descrizione degli stati conformazionali delle IDPs, un problema che può essere mitigato dallo sviluppo di campi di forza ad hoc per tali proteine. Nonostante queste limitazioni, abbiamo fornito una prima descrizione degli stati conformazionali della terza UIM di AT3 mostrando che ha una bassa propensione a popolare conformazioni ad α elica in soluzione. Abbiamo analizzato le caratteristiche strutturali della regione C-terminale di AT3 3UIM e fornito risultati per capire la sua scarsa propensione elicoidale. I risultati ottenuti suggeriscono che la posizione di residui subottimali per la geometria delle eliche in 3UIM hanno un ruolo nella sua scarsa propensione a popolare conformazioni elicali stabili in soluzione. Abbiamo inoltre individuato nell'ensemble di AT3 3UIM libero, conformazioni che ricordano la struttura NMR di una UIM canonica in complesso con l'ubiquitina. A causa della bassa propensione per gli stati elicali in soluzione, sono necessari ulteriori studi per chiarire se la terza UIM non è una UIM canonica e quindi possa legarsi con affinità molto bassa o non legarsi con l'ubiquitina, o se un *ensemble* disordinato possa essere una caratteristica comune anche ad altre UIM.

Nel capitolo 3, è presentato l'altro soggetto principale di questa tesi, vale a dire il ruolo di cofattori metallici e degli effetti indotti da metalli non essenziali sulla struttura e la dinamica di domini di fattori di trascrizione, come ad esempio il dominio *zinc finger* della proteina SUPERMAN e il dominio di legame al DNA (DBD) di p53. P53 svolge un ruolo centrale nei processi cellulari, attraverso l'interazione con molteplici interattori, ma attualmente vi è una conoscenza ancora molto limitata degli effetti indotti sulla struttura e funzione di p53 a seguito del legame con altre biomolecole, come il DNA. Nel sottocapitolo 3.1, presentiamo lo studio MD che abbiamo svolto su p53 con lo scopo di identificare le regioni distali dal sito di legame al DNA che sono allostericamente modulate dall'interazione con il DNA stesso e possono regolare la capacità di p53 di reclutare altri cofattori biologici. Abbiamo impiegato sia simulazioni MD canoniche che metodi di *enanched-sampling* e analizzato i risultati impiegando metodi ispirati alla teoria dei grafi per studiare il dominio DBD di p53 in assenza e presenza di DNA. Abbiamo descritto cambiamenti conformazionali e *paths* intramolecolari di comunicazione a lungo raggio in p53 DBD a seguito dell'interazione con il DNA. Abbiamo mostrato che il legame al DNA promuove un cambiamento conformazionale nel loop S6-S7, una regione a 3 nm di distanza dal sito di legame al DNA di p53, aumentando di più di quattro volte la popolazione di uno stato minore, che è presente anche nella proteina libera in soluzione. Abbiamo identificato un'interfaccia di p53 che lega partner biologici di p53 connessi a funzioni trascrizione-indipendenti che non è più accessibile negli stati indotti da interazioni con il DNA. Inoltre, abbiamo identificato 7 partner che è probabile interagiscano con p53 DBD in prossimità del loop S6-S7 e possano essere regolati dai cambiamenti allosterici indotti dal DNA. Un esempio di questi interattori è Ku70 che è coinvolto nella segnalazione apoptotica. I nostri risultati suggeriscono un possibile meccanismo di protezione di p53 da interazioni con partner che svolgono la segnalazione apoptotica trascrizione-indipendente quando p53 è legato al DNA e deve svolgere le sue funzioni di regolazione della trascrizione.

Nel sottocapitolo 3.2 ci siamo concentrati sullo studio degli effetti indotti dal cadmio, un noto metallo inquinante e cancerogeno, sulle proteine che legano lo zinco. La caratterizzazione strutturale delle zinco-proteine e delle alterazioni associate alla tossicità del cadmio sono essenziali per comprendere il ruolo dei metalli in questi problemi biologici altamente rilevanti. In questo contesto, in quanto approfondite analisi sperimentale sono ancora mancanti, gli approcci computazionali possono essere di grande valore per lo studio strutturale delle zinco-proteine, ma nei campi di forza attuali sono presenti approssimazioni sia nella descrizione della coordinazione con i metalli che nella modellazione del cadmio. Abbiamo quindi sviluppato un protocollo basato sulla meccanica molecolare classica (MM) e su calcoli di chimica quantistica (QM) per derivare parametri ottimizzati nel campo di forza per ioni metallici, che possono poi essere utilizzate in simulazioni MD. Il protocollo da noi sviluppato ha lo scopo di descrivere in modo efficace la coordinazione con lo zinco di p53 e di altre metallo-proteine e di indagare a fondo gli effetti strutturali indotti dal cadmio. In primo luogo, abbiamo fornito una descrizione del protocollo e delle procedure utilizzate per ottimizzare i parametri del campo di forza utilizzando un approccio di *simulated annealing* per riprodurre con i calcoli MM la geometria di minima energia ottenuta mediante QM, e in aggiunta riprodurre le curve di dissociazione del legame. Abbiamo applicato il nostro protocollo per ottenere parametri ottimizzati per il sito di coordinazione del dominio *zinc finger* che è composto da due cisteine e due istidine (Zn-Cys$_2$His$_2$). Abbiamo testato il comportamento dei nostri parametri ottimizzati in simulazioni MD del dominio *zinc finger* della proteina SUPERMAN confrontandole con simulazioni MD effettuate con il campo di forza CHARMM22/CMAP. Abbiamo mostrato l'efficacia del nostro approccio e dei parametri ottimizzati che danno risultati in buon accordo con la struttura sperimentale e i calcoli QM. Abbiamo anche sviluppato nuovi parametri per il cadmio per la geometria di coordinazione Cys_2His_2. Abbiamo dimostrato che il nostro primo set da risultati migliori rispetto ai parametri attualmente presenti in letteratura, mostrando alterazioni conformazionali simili a quelle osservate sperimentalmente, ma sono in corso ulteriori ottimizzazioni per descrivere accuratamente i meccanismi molecolari di tossicità del cadmio.

La tesi si chiude nel capitolo 5, in cui le conclusioni sul progetto sono discusse sottolineando che simulazioni MD eseguite con metodi e campi di forza allo stato dell’arte, e se integrate con tecniche biofisiche sperimentali, sono uno strumento potente ed efficace per lo studio della dinamica delle proteine e la loro relazione con la struttura e le funzioni biologiche.

Publications

This thesis has led to the following publications and two other manuscripts presently in preparation:

Subchapter 2.2

Intramolecular interactions stabilizing compact conformations of the intrinsically disordered kinase-inhibitor domain of Sic1: a molecular dynamics investigation.

Lambrughi M., Papaleo E., Testa L., Brocca S., De Gioia L. and Grandori R. (2012)

Front. Physiol., 3:435.

Subchapter 2.3

The conformational ensemble of the disordered and aggregation-protective 182-291 region of ataxin-3.

Invernizzi, G., Lambrughi, M., Regonesi, M.E., Tortora, P., Papaleo, E. (2013).

Biochim. Biophys. Acta, 1830 (11):5236-47

Subchapter 3.2

DNA-binding protects p53 from interactions with cofactors involved in transcription-independent functions.

Matteo Lambrughi, Luca De Gioia, Francesco Luigi Gervasio, Kresten Lindorff-Larsen, Ruth Nussinov, Chiara Urani, Maurizio Bruschi, Elena Papaleo

Submitted.

Related publications:

A comparative study of Whi5 and retinoblastoma proteins: from sequence and structure analysis to intracellular networks.

Hasan MM, Brocca S, Sacco E, Spinelli M, Papaleo E, Lambrughi M, Alberghina L, Vanoni M. (2012)
Front. Physiol., 4:315

PyInteraph: a framework for the analysis of interaction networks in structural ensembles of proteins.

Tiberti M., Invernizzi G., Lambrughi M., Inbar Y., Schreiber G., and Papaleo E. (2014)
J Chem Inf Model, 54: 1537-51.

Communication routes in ARID domains between distal residues in helix 5 and the DNA-binding loops.

Invernizzi G., Tiberti M., Lambrughi M., Lindorff-Larsen K., Papaleo E. (2014)
PLoS Comput. Biol. 2014, 4:e1003744.

1 Introduction

1.1 Protein dynamics

Proteins are essential components of cell that have been demonstrated to be highly dynamic systems (Vendruscolo et al. 2007, Zhuravlev and Papoian, 2010). Protein dynamics occur in an array of conformational transitions ranging from small and local fluctuations in atomic positions occurring near native state, to collective movements of entire domains, subunits and large-scale changes through the whole structure of the protein. Nevertheless, the most general fluctuations are usually small in magnitude, not exceeding several Ångstroms, and lie in the nanoseconds (ns)-sub-ns frequency range. It is now well established that the biological functions of proteins can be modulated or determined by their dynamics and these are intimately linked to the three-dimensional (3D) structure (Karplus et al. 2005, Vendruscolo et al. 2007, Henzler-Wildman and Kern, 2007). Therefore not only the description of protein structure but also the knowledge of their dynamics is crucial to understand their biological functions and has a practical return in biotechnological and pharmacological area e.g. in the design of new drugs (Ozbabacan et al. 2010, Teilum et al. 2011). In the last years several questions related to protein dynamics, such as the mechanism of allosteric changes, have been intensively studied (Fischer et al., 1894, Koshland et al. 1958, Straub and Szabolcsi, 1964, Závodszky et al., 1966, Tsai et al., 1999, Popovych et al., 2006, Goodey and Benkovic, 2008, Klepeis et al. 2009, Whitley and Lee, 2009, Kar et al., 2009, Manley and Loria, 2012, Nussinov and Tsai, 2013, Tsai and Nussinov, 2014), but have not been completely elucidated yet. In this context, a detailed atomistic description of the dynamical fingerprint in proteins is particularly needed, considering its key role in different aspects of protein function, including thermal stability, catalysis macromolecular recognition, allosteric regulation and maintenance of the three-dimensional architecture (K. Henzler-Wildman et al., 2007, Del Sol et al. 2009, Hilser 2010, Ma and Nussinov, 2010). In particular it is of fundamental importance to get knowledge of the free energy landscape of the proteins native state and describe even their less populated states (Baldwin and Kay, 2009). Indeed proteins also in their free and unaltered state can undergo conformational alterations and they exist not as a single conformation but as an ensemble of numerous states (Montlagh et al., 2012, Frauenfelder et al., 1991). These can be viewed as conformations near their native state, which are in dynamical equilibrium (Lindorff-Larsen et al., 2005, Vendruscolo et al., 2007, Ozbabacan et al., 2010) and interconvert by fluctuations on the picosecond (ps) or ns timescale. These conformations are often only transiently populated in comparison to the major state of the protein but it is essential to describe them in details, since they can resemble states relevant for the biological functions such as conformations suitable to interact with a ligand or catalytically-competent states. These ensembles of pre-existing populations can be affected by interactors, post-translational modifications and mutations leading to an equilibrium shift of conformational states and dynamics (Gunasekaran et al., 2004, Del Sol et al. 2009). This process of modulation of populations not necessarily involves a change of protein shape and structure (Tsai et al., 2008) but can rely only on changes in dynamics (Popovych et al., 2006, Tsai and Nussinov, 2014). For example it has been proposed that an initial subtle perturbations in the network of residue contacts and dynamics in a localized site can result in perturbation of communication paths between residues across the protein structure and define long range induced effects (Cui et al., 2008).

In this context, the recent advances in nuclear magnetic resonance (NMR) spectroscopy (Esteban-Martin et al., 2010) and molecular simulations approaches (Dodson, et al., 2008.) made it possible to analyse conformational ensembles of proteins and give description of their dynamics and extract information on functional properties. Especially computational techniques and molecular simulations are becoming an essential tool in the study of protein structure and dynamics. In the past years computational methods based on first-principles physics has been developed to model, study and predict structure and dynamics of proteins at atomic-level, permitting to investigate the features of chemical systems like thermodynamic and functional properties (Lindorff-Larsen et al., 2005, Klepeis et al., 2009, Dror et al., 2012). Nowadays *in silico* simulations of molecular systems are essential tools in several research area, since they can provide information and details on biological mechanisms at a time and spatial scales difficult to investigate with only experiments (Shaw et al., 2012). In fact thanks to the recent improvements in both hardware (Shaw et al. 2014) and software (Harvey et al., 2009) for simulations performance, the computational cost, that biomolecular simulations requires has been progressively reduced so that now larger timescales are accessible. Moreover, the community is doing remarkable efforts to improve the physical models (i.e. the force fields) to describe the protein and the solvent in the simulations to provide a more accurate description of protein dynamics (Lindorff-Larsen et al., 2012, Lindorff-Larsen et al., 2011, Papaleo et al., 2014; Palazzesi et al., 2014). Simulations can provide atomic-level description till milliseconds timescales of biomolecules and their complexes. Moreover they turned out to be a successful technique to study function−structure−dynamics relationships in protein and investigate essential biochemical events like protein allostery, conformational changes associated with function, molecular recognition, folding and drug binding (Dror et al., 2012). Furthermore these techniques integrated to graph theory permit the description of communication paths of residues by the networks of intramolecular interactions, in a dynamic perspective. Moreover molecular simulations can provide relevant information on the mechanistic and functional aspects of protein system (Bahar et al. 2010), thanks for example to the analysis of the cross-correlations of atomic fluctuations (Hunenberger et al., 1995) or long range pathway of communicating residues (Morra et al., 2009, Vishveshwara et al., 2009). Simulations can be helpful to compare and integrate experimental data and efficiently contribute to rationalize them in particular if simulations and experiments work in a continuous cross-talk (Fenwick et al., 2011, Salvatella et al. 2011, Papaleo et al. 2012). Several experimental techniques have been used to effectively describe and characterize dynamics of proteins but they generally provide ensemble averaged information, partially missing the contribution coming from individual populations (Dror et al., 2012). In particular NMR spectroscopy is a powerful tool in the experimental analysis of macromolecular dynamics, because it provides information about both structure and dynamics at atomic resolution that can be used to validate the predictions of computational simulations (Best and Vendruscolo 2004, Clore and Schwieters 2004b, Lindorff-Larsen et al., 2005, De Simone et al., 2009).

1.2 Computational methods

1.2.1 Molecular Dynamics

Different computational approaches have been developed based on the different level of first-principle physics and physical models used and the most rigorous approach is based on Quantum Mechanics (QM). QM methods are based on the description of the behavior of molecules at subatomic level with the relativistic Schrodinger equation but the direct solution of this equation is too computationally demanding and infeasible for the study of biomolecules. To treat large systems like biological macromolecules and proteins the standard method is a technique known as all-atom molecular dynamics (MD) simulations, in which the system is treated by the laws of classical physics. In a system composed by fixed particles the potential energy depends on the interactions between the atoms thanks to their reciprocal positions and the potential energy varies at the variation of the geometry of the system. In this methodology the atoms of molecular system are treated as classical particles whose interactions are described by potential energy functions that approximate the real interactions between particles. The potential energy of the atoms configuration is represented by a mathematical model based on the classical mechanics, called force field. Usually force fields and their parameters are designed based on first-principles physics and fitting to quantum mechanical calculations and experimental data. In the MD, through the functions of the force field, the molecule is represented as an ensemble of spheres with mass, representing the atoms, linked by springs that represent the interactions treated with harmonic potentials and the system is processed by the law of classical physics. Although MD simulations don’t exactly model the physics of the system and have simplifications, such as the non explicit treatment of the electrons that do not permit to represent process that involves the formation or the breaking of covalent interactions, it can provide sufficient close approximation to effectively treat large and complex molecular systems with thousands of atoms, as proteins, with affordable computational cost and capture a wide range of critical biochemical processes with very high level of reliability. The choice of parameters, or the parameterization of the force field, depends on the type of atoms involved and the context in which the molecular interaction takes place, and since it is impossible to consider all possible interactions found in nature, the force field are parameterized in an adequate manner to describe precise molecular systems such as proteins.

The functions of the potential comprised in a force field are divided usually in terms or binding interactions, and non-bonding. The binding interactions are defined by the equation:

$$\begin{aligned}
&v_{bonded}(r) = \sum_{bonds}\frac{1}{2}k^{r}_{ij}(r_{ij}-r_{eq})^{2} + \\
&+ \sum_{angles}\frac{1}{2}k^{\theta}_{ijl}(\theta_{ijl}-\theta_{eq})^{2} + \\
&+ \sum_{torsions}\frac{1}{2}k^{\phi}_{ijlm}(\phi_{ijlm}-\phi_{eq})^{2}
\end{aligned}$$

The terms of the equation express respectively the component relating to the stretching of the bonds, the variation of the angles and the rotation of the bonds and they require respectively 2, 3 and 4 atoms to be calculated. The first term in the equation expresses the component of the potential energy related to the interaction between two atoms forming a covalent bond and describes the variation in function of the stretching of the bond itself, r_{ij} is the distance between atoms *i* and *j*. The second term describes the variation of potential energy associated with the variation of the bond angles, and q_{ijl} is the amplitude of the angle formed by the atoms *i*, *j*, *l*. The third term represents the change in potential energy associated with the rotation of the angles and depends on the variation of the dihedral angle formed by four consecutive atoms, and ϕ_{ijlm} is the torsional angle between the atoms *i, j, l,* and *m*. These potential for binding interactions are modelled as harmonic oscillators and the values at equilibrium, in conditions of minimum energy, r_{eq}, θ_{eq}, ϕ_{eq} and force constants k are defined as parameters in the force field and depend on the interactions considered.

The non-bonding potential describes the interactions between pairs of atoms not directly linked to each other and separated by at least three bonds (ixxj). Non-bonded interactions comprise a Lennard-Jones potential to describe the interactions of Van der Waals and a function for calculating the Coulomb electrostatic interactions:

$$v_{non-bonded}(r) = \sum_{couples\{i,j\}} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \\ + \sum_{couples\{i,j\}} \frac{Q_i Q_j}{4\pi\varepsilon_0 r_{ij}}$$

In the Lennard-Jones potential σ is the constant related to the atomic radius, while for in the Coulomb potential Q_i and Q_j are the charges of the two atoms involved, r_{ij} is the distance and ε_0 is the dielectric constant in vacuum. The parameters σ, ε are defined in the force field. The parameterization of the force field, that is the determination of the constants and parameters present in his equations, is made empirically based on experimental data or computational data obtained by quantum-mechanics. The force fields developed are many such GROMOS, AMBER, ECEPP, CHARMM and a number of improved force field have been introduced in the past years after extensive validation against experimental data (Lindorff-Larsen et al., 2010, Lindorff-Larsen et al., 2012).

1.2.2 The Statistical Mechanics

The molecular dynamics (MD) simulations allow to obtain information at the microscopic level, such as the position of atoms and their velocities, and the use of this information to derive high-order and macroscopic properties of the system in analysis such as for example energy changes and mechanisms associated with structural alterations. The connection between the microscopic details of the simulations with the macroscopic properties of biological systems is done by statistical mechanics that is a branch of physical science that studies macroscopic systems from a molecular point of view. The final aim is to predict macroscopic phenomena from the properties of the individual elements that make up the system. It has been demonstrated that the molecular systems are generally characterized by a limited number of parameters that describe their global properties, such as temperature or pressure, or the number of particles, collectively called macrostates or thermodynamic states. Moreover, a molecular system can be defined at the microscopic level by a set of atomic positions and velocities of its atoms. Each individual set of possible positions and associated velocities for the system is defined as a microstate. The positions and velocities of the atoms used as coordinates allow to define a multidimensional space defined as phase space, in which the points represent univocally all and only the possible states (microstates) or conformations of the system. An ensemble is a collection of points in phase space that satisfy a particular thermodynamic state or macrostate. Therefore, ensembles can be analysed as NVT (canonical ensemble) or NPT (isobaric-isotermal ensemble) where the macrostate is characterized by a fixed number of N atoms, a fixed pressure P and constant temperature T. In this way, the macroscopic quantities of a system are represented by the values in the ensemble. The ensemble can be described as a population of states of the system that is distributed in a stochastic way compared to a mean value, which can be approximated to the value observed experimentally. Through sampling methods, such as molecular dynamics, which generate an ensemble composed by a set of microstates (positions and velocities) in function of time (the trajectory of the system), it is possible to derive macroscopic quantities, as the free energy of system, from the average of the values that form the ensemble. This approach assumes that the average of the values of the ensemble coincides with the average of the values assumed by the system in its temporal evolution during the sampling of the phase space. So only a sampling that last for infinite time would end up generating an ensemble that comprises all the states belonging to the phase space. MD simulations has to be conducted long enough to get a good sampling that allows to obtain real values for the macro features of the system as well as maintaining feasible computational cost.

1.2.3 Molecular Dynamics

Although it is impossible to sample all points of the phase space in an affordable time, the MD can sample the phase space of the system in an efficient way, permitting to obtaining detailed information at the microscopic and the macroscopic level and on dynamics at different but still limited timescales (Dror et al., 2012). The MD is based on the use of Newton's second law of motion and the integration with the equations of motion. The force exerted on the system can be calculated by the partial derivatives of the energy, obtained from the force field, in relation to the coordinates of the system. The forces on each atom permit to compute accelerations and velocities, extrapolate the new positions of the atoms and determine the configuration of the molecular system after a certain period of time. This process if iteratively repeated for each time interval allows to generate a trajectory that describes the positions, velocities and accelerations of the particles as a function of how they vary during time and allows to calculate the properties of the molecular system from the sampled ensemble. The execution of a MD simulation of a molecular system requires the initial position of the atoms, which is obtained from high resolution experimental structures, such as by NMR or by X-ray crystallography, and the initial velocities of the atoms, which are randomly defined by a Maxwell Boltzmann distribution at a certain temperature, in order to assign the correct level of kinetic energy for the temperature of simulation. The MD is now one of the more accurate theoretical technique to carry out studies of protein systems. This computational method allows to study the behavior of a molecular system and MD simulations have yielded detailed information on fluctuations and conformational changes of proteins and macromolecules. This method is frequently used to study the structure, dynamics and thermodynamics of biological molecules and their complexes. Nowadays with the advent of new technologies MD simulations are used to study increasingly complex systems, to characterize and refine the structures derived from experiments of X-ray crystallography and NMR, to investigate the phenomena of folding / unfolding, to study the docking interactions between biological molecules and with drugs and also in the study of the mechanisms involved in mutations (Dror et al. 2012). In the present project we used multi-replicate full atom explicit solvent MD simulations showing that, in combination with improved force field and integrated with structural and dynamical data from experimental techniques, they are effective in the characterization of protein structure, dynamics and functions also in complex biological systems and important for the implications on human health, like intrinsically disordered proteins and metal binding proteins. The accuracy of MD results strictly depends on the force field used and it is known that despite the recent improvements (Lindorff-Larsen et al., 2012, Best et al., 2012) force fields are still not perfects and cannot simulate with accuracy several properties of proteins. For example the currently used force fields have the risk to sample overcompact states of disordered domain and they miss the explicit treatment of polarization effects in the interaction between atoms and have propensity to overestimate salt bridges. To overcome those MD limitations and avoid misinterpretation of the results we collected experimental data from different biophysical techniques that have been integrated with computational data. Moreover canonical MD, despite the recent advances in the computation performance of the high parallelized computer chips and new special purpose architectures, like graphics processing units (GPUs) and optimized software, is still not suitable for some tasks in the study of proteins, such as the accurate study of the free energy landscape or the study of conformational ensembles of highly heterogeneous proteins. In fact in the classical MD simulations the system has the risk to be trapped in a local energy minimum for long time, providing a limited sampling of the conformational sample of the protein (Lindorff-Larsen et al., 2012). In these cases, we used enhanced sampling simulations, a variety of algorithms that can be applied in combination with MD simulations to increase the sampling of the conformational space accessible to the protein under investigation. Those methods have the potential to sample conformational changes occurring on longer timescales that remain inaccessible to all-atom canonical MD and also to re-sample different states of the protein in solutions to allow for statistically significant estimation of their population (Sugita and Okamoto, 1999, Sutto et al., 2012). In particular, we employed Replica Exchange MD (REMD), in which a number of different copies (replicas) of the system are simulated in parallel at different temperatures and exchanges of configurations are tried periodically between pairs of replicas (Sugita and Okamoto, 1999). The advantage of this method is that if the trajectory is temporarily trapped in a local minimum can exchange with a higher-temperature replica and it can thus, more easily, cross high-energy barriers and allow the system to reach equilibrium more quickly (Allison and van Gunsteren, 2009, Markwick et al., 2009).

We studied MD ensemble giving details of structural properties, insights into the paths of long-range structural communication and interaction networks of proteins, providing characterization of dynamics in order to elucidate their relevance in protein biological functions. We analysed our MD simulations with different tools to estimate protein flexibility, identify clusters of similar conformations in the MD trajectories, and to describe the networks of intramolecular interactions and paths of communications (Ghosh and Vishveshwara, 2007, Papaleo et al., 2012). We employed methods as Principal Component Analysis (PCA) (Amadei et al., 1993; van Aalten et al., 1997) and Protein Structure Network (PSN) approaches integrated to evaluation of correlated motions in the ensemble by the Linear Mutual Information (LMI) metric (Lange and Grubmuller 2006, Angelova et al. 2011).

1.3 Methods for the analysis of MD ensemble

1.3.1 Principal component analysis

Principal Component Analysis (PCA), also known as Essential Dynamics (ED), is a mathematical method that permits a detailed and comprehensive analysis of protein dynamics signature, revealed by high-amplitude concerted motion in MD trajectories (Amadei et al. 1993, Amadei et al. 1999). In fact, proteins are characterized by both very fast vibrational motions and larger amplitude motions with a low frequency, which are, at least in several cases, the motions of interest when structure-function relationship are investigated. After superposition to a common reference structure of the conformations sampled during MD simulations, PCA is applied by construct the mass-weighted covariance matrix of atomic positional fluctuations of a subset of atoms of interest. The symmetric matrix is diagonalized by an orthogonal coordinate transformation, which contains the eigenvectors, or principal components, of the covariance matrix. The eigenvalues obtained correspond to the mean square eigenvector coordinate variations and contain the contribution of each principal component to the total fluctuations in protein. The principal components are determined in the way that the first component describes the largest part of the data variability in the dataset, and so also for the subsequent components with the remaining variability. In this way PCA permits to analyse a multidimensional dataset, like the structural ensemble sampled during MD simulation, converting it into a new system of unrelated variables in which the first principal components are the coordinates. They are used to obtain a new accurate representation of the data filtered of fewer dimensions, while maintaining most of the information and the coordinates that express larger fluctuations and account for most of the variability of the system. It is also possible to project on the three-dimensional structure of the protein the regions where the motions described by the first eigenvectors are located to link protein dynamics to structure and identify the directions of maximum variance. Furthermore, the PCA analysis was used to evaluate the quality of the sampling achieved in a MD simulation by analysing the similarity between principal components of protein motions on random diffusion (Berk 2000), calculating the cosine content of the principal components. This is an index calculated from the covariance matrix that shows the similarity between the projections of the trajectory on the principal components with the cosine function. Indeed it has been demonstrated that an insufficient sampling could lead to high value of cosine, that is representative of random motions.

1.3.2 Network theory in biology

Recently, the use of network theory to describe biological phenomena is becoming popular, accounting for applications in different fields from ecosystem modelling, structural and system biology. Network theory, thanks to the increasing computational resources available, provides tools and algorithms to describe and analyse large data sets of complex systems. The networks, as postulated by graph theory, consist of a set of interacting elements (nodes or vertices), which are connected in pairs by links (contacts, edges, interactions) (Barabási and Oltvai, 2004, Böde et al. 2007, Csermeley et al. 2012, Csermely et al. 2013, Lovász, 2012). In the case of molecular networks of biological system, nodes can be amino acids, proteins or other macromolecules, depending on the field of the application. Network edges consist of the connections between the nodes and represent the relationship that intercourse in the network. For example, in the molecular networks of biological systems edges represent physical or functional interactions of two network nodes (Zlatic et al., 2009). In a protein structure network, instead, the edges can describe the intramolecular interactions between residues, such as hydrophobic and electrostatic interactions or hydrogen bonds, or even more simply the Van der Waals contacts. Edges usually are associated with a weight that represents the intensity (strength, probability, affinity) of the interaction. The networks that have been found so far in protein science feature small worlds properties. A small world network is a network in which two elements of the network are separated by only a few other elements. Important nodes of this class of networks are the so-called hubs, i.e. elements which feature a higher number of connections/edges with respect to other nodes of the network. Hubs are believed to play a central role in the network architecture and in the communication between the different regions of the network. In a network, we can also identify other smaller components and they can be divided in modules (communities, groups), which often form a hierarchical structure.

1.3.3 Protein structure networks

The network paradigm has been extensively used to describe the structure, topology, and dynamics of proteins (Csermely et al., 2013). In particular the communication paths and intramolecular interactions between residues in a protein can be collectively represented in the form of a network, namely Protein Structure Network (PSN). This method is based on the fact that structural effects can be transmitted at distal sites through communication paths involving contacts between residues. The application of such approach is very important since it permits to investigate long-range communication and intramolecular interaction networks, that are known to be essential in determining the protein structure (Brinda et al., 2005, Böde et al., 2007, Scarabelli et al., 2010, Atilgain et al., 2012).

In PSN protein structures are represented as networks (graphs) where amino acid residues are the nodes and their interactions are the edges (Böde et al., 2007). In most protein structure network representations, the nodes are usually the amino acid side chains although occasionally PSN nodes are defined as the atoms of the protein. The side-chain representation is justified by the assumption that side-chains atoms have concerted movements. Edges of PSN are defined using the distance between amino acid or atoms and can be described with different strategies, as for example distances between Cα atoms, the center of mass of each side chains, atomic contacts, van der Waals interactions, etc. Edges of PSN connect amino acids having a distance below a cut-off distance, which is usually between 4 and 8.5 Å (Artymiuk et al., 1990, Vishveshwara et al., 2009, Doncheva et al., 2011, Doncheva et al., 2012, Csermely et al., 2012). Most PSN edges are not associated to a weight, and only very recently the application of PSN methods to the analysis of proteins has made use of weights, which can be the persistence of the interaction in the ensemble or the correlations between each pair of nodes during dynamics. PSN have small worlds properties (Atilgan et al., 2004, Vendruscolo et al., 2002), a crucial feature for the fast transmission of information in protein structures from one distal site to the sites where a conformational change need to occur. Indeed, in the small-world of PSN amino acids can communicate through the shortest paths available and generally residues communicate with each other by only a few steps. Moreover conformational changes in protein structure during dynamics are transmitted through multiple paths with often several nodes in common (del Sol et al. 2009). These type of networks are usually characterized by the presence of small number of central nodes that make a large number of interactions with the other nodes in the network, called hubs of a PSN. Hubs are believed to play an important role in protein structures and have been suggested important for the thermodynamic stability of the protein, for the formation of secondary structure elements, for allosteric mechanisms or even catalysis, substrate and co-factor binding (Böde et al., 2007) even if none clear validation of these statements have been provided so far (Kannan and Vishveshwara, 1999, Brinda and Vishveshwara, 2005, Konrat et al., 2009, Csermely et al., 2012). Some studies also suggest that hubs may be the residues of the protein that are less tolerant to mutations. Moreover PSN approaches may help to identify key amino acids involved in structural communication between distal sites of a protein. In particular, graph theory provides different algorithms to calculate paths of communications between nodes in the network and those paths in PSN might be important for long-range communication or even for allosteric mechanisms (Chennubhotla and Bahar, 2006, Ghosh and Vishveshwara, 2008, Tehver et al., 2009, Vishveshwara et al., 2009, Csermely et al., 2012). Despite the great potential of these approaches, lot is still unknown in networks analysis of proteins since it is difficult to obtain predictions from these analyses and to know exactly if the calculated paths are relevant for the functional properties.

Recently, several methods have been developed to calculate the PSN from protein structural ensembles, as the ones obtained by MD simulations, and to then analyse the networks and detect hubs and paths of communications. These PSN approach relies on the integration of metrics that estimate dynamically-coupled residues, which is based on the description of the atomic contacts between residues that also feature coupled motions from MD ensemble (Ghosh and Vishveshwara, 2008, Seeber et al., 2011, Papaleo et al., 2012). Several metrics have been proposed to estimate the coupled motions from the MD ensemble and used with PSN calculations, but they often have problems for the convergence of the analysis and statistical significance of the results. We here employed Linear Mutual Information (LMI) as a metrics to evaluate correlated motions (Lange and Grubmüller, 2006, Raimondi et al., 2008) since, together with Dynamical Cross-correlation matrix (DCCM) approach (Ghosh and Vishveshwara, 2008), are proposed to be accurate methods to study MD ensembles. LMI unlike cross-correlation has the advantage of not depending on the relative orientations of the fluctuations, permitting to identify correlated motions despite the difference between their orientations in space. In the used PSN approach each residue is considered as a graph node and the network of interacting residues in the protein is defined from the number of non-covalently intramolecular interactions, using a calculated interaction strength value (I_{ij}) as edge weight, where i and j are the residue. The interaction strength value is calculated:

$$I_{ij} = \frac{nij}{\sqrt{N_i N_j}}$$

where n_{ij} is the number of distinct atom pairs between residues *i* and *j* within a distance cutoff of 4.5 Å. *Ni* and *Nj* are normalization values for residues *i* and *j* which account for the different interaction propensity due to larger or smaller side-chains and are obtained from a statistically significant protein dataset. A PSN edge is retained between two nodes only if its interaction strength is higher than a cutoff value I_{min}. The residues involved in more than four edges are referred as hubs at a specific $I_{min.}$ Moreover clusters can be identified as a set of connected residues in the graph. Each node is assigned to a cluster if it establishes a link with at least one node already in the cluster. If no interactions are present the node is assigned to another cluster. The cluster size is defined by the number of its nodes at a specific I_{min} and the so-called I_{crit} can be calculated, that is the value at which the size of the largest cluster in the PSN graph significantly changes (Brinda and Vishveshwara, 2005). Here I_{min} value was considered equal to I_{crit} since it generally permits to well discharge weak interactions. Several tools and software have also made available to calculate PSNs and their properties. PSN approaches are for examples implemented in the Wordom toolkit for MD analysis (Seeber et al., 2011), which have been used in this thesis. These approaches consider the atomic contacts and permit to describe the van der Waals effects, that are crucial for long range communications, but they do not usually take into account the different chemical-physical of the residues. Thus, in our group, we also developed a suite of tools, called PyInteraph (Tiberti et al., 2014), to calculate networks of weak intramolecular interactions from conformational ensembles of proteins, to complement the classical PSN information. We used these methods for most of the target proteins discussed in this thesis.

1.4 References.

Aftabuddin, M. and S. Kundu (2007). Hydrophobic, hydrophilic, and charged amino acid networks within protein. *Biophys J* 93(1): 225-231.

Allison, J. R. and W. F. van Gunsteren (2009). A method to explore protein side chain conformational variability using experimental data. *Chemphyschem* 10(18): 3213-3228.

Alves, N. A. (2007). Unveiling community structures in weighted networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 76(3 Pt 2): 036101.

Amadei, A., Linssen, A. B., and Berendsen, H. J. (1993). Essential dynamics of proteins. *Proteins* 17, 412-425

Amadei, A., Ceruso, M. A., and Nola, A. D. (1999). On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. *Proteins* 36, 419-424.

Amitai, G., A. Shemesh, E. Sitbon, M. Shklar, D. Netanely, I. Venger and S. Pietrokovski (2004). Network analysis of protein structures identifies functional residues. *J Mol Biol* 344(4): 1135-1146.

Arrigoni, A., B. Grillo, A. Vitriolo, L. De Gioia and E. Papaleo (2012). C-terminal acidic domain of ubiquitin-conjugating enzymes: A multi-functional conserved intrinsically disordered domain in family 3 of E2 enzymes. *J Struct Biol*.

Artymiuk, P. J., D. W. Rice, E. M. Mitchell and P. Willett (1990). Structural resemblance between the families of bacterial signal-transduction proteins and of G proteins revealed by graph theoretical techniques. *Protein Eng* 4(1): 39-43.

Atilgan, A. R., P. Akan and C. Baysal (2004). Small-world communication of residues and significance for protein dynamics. *Biophys J* 86(1 Pt 1): 85-91.

Atilgan, C., O. B. Okan and A. R. Atilgan (2012). Network-based models as tools hinting at nonevident protein functionality. *Annu Rev Biophys* 41: 205-225.

Babu, C. S. and C. Lim (2006). Empirical force fields for biologically active divalent metal cations in water. *J Phys Chem A* 110(2): 691-699.

Bahar, I., T. R. Lezon, L. W. Yang and E. Eyal (2010). Global dynamics of proteins: bridging between structure and function. *Annu Rev Biophys* 39: 23-42.

Barabasi, A. L. and Z. N. Oltvai (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5(2): 101-113.

Barberis, M. (2012). Sic1 as a timer of Clb cyclin waves in the yeast cell cycle - design principle of not just an inhibitor. *FEBS J*.

Barberis, M., C. Linke, M. A. Adrover, A. Gonzalez-Novo, H. Lehrach, S. Krobitsch, F. Posas and E. Klipp (2012). Sic1 plays a role in timing and oscillatory behaviour of B-type cyclins. *Biotechnol Adv* 30(1): 108-130

Böde, C., Kovács, I.A., Szalayb S.M., Palotaib, R., Korcsmáros,T., Csermely, P., (2007) Network analysis of protein dynamics. *FEBS Letters* 581:2776-2782.

Brinda, K.V., Vishveshwara, S. (2005). A network representation of protein structures: implications for protein stability. *Biophys J*. 89(6):4159-70.

Hess, B. (2000). Similarities between principal components of protein dynamics and random diffusion. *Phys. Rev.*, 62, 8438

Best, R. B. and M. Vendruscolo (2004). Determination of protein structures consistent with NMR order parameters. *J Am Chem Soc* 126(26): 8090-8091.

Boehr, D. D., R. Nussinov and P. E. Wright (2009). The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol* 5(11): 789-796.

Breydo, L. and V. N. Uversky (2011). Role of metal ions in aggregation of intrinsically disordered proteins in neurodegenerative diseases. *Metallomics* 3(11): 1163-1180.

Brocca, L., J. Cannavino, L. Coletto, G. Biolo, M. Sandri, R. Bottinelli and M. A. Pellegrino (2012). The time course of the adaptations of human muscle proteome to bed rest and the underlying mechanisms. *J Physiol* 590(Pt 20): 5211-5230.

Brocca, S., L. Testa, F. Sobott, M. Samalikova, A. Natalello, E. Papaleo, M. Lotti, L. De Gioia, S. M. Doglia, L. Alberghina and R. Grandori (2011). Compaction properties of an intrinsically disordered protein: Sic1 and its kinase-inhibitor domain. *Biophys J* 100(9): 2243-2252.

Bucciantini, M., E. Giannoni, F. Chiti, F. Baroni, L. Formigli, J. Zurdo, N. Taddei, G. Ramponi, C. M. Dobson and M. Stefani (2002). Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* 416(6880): 507-511.

Chennubhotla, C. and I. Bahar (2006). Markov propagation of allosteric effects in biomolecular systems: application to GroEL-GroES. *Mol Syst Biol* 2: 36.

Chennubhotla, C., Z. Yang and I. Bahar (2008). Coupling between global dynamics and signal transduction pathways: a mechanism of allostery for chaperonin GroEL. *Mol Biosyst* 4(4): 287-292.

Chiti, F. and C. M. Dobson (2006). Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem* 75: 333-366.

Chiti, F., N. Taddei, M. Bucciantini, P. White, G. Ramponi and C. M. Dobson (2000). Mutational analysis of the propensity for amyloid formation by a globular protein. *EMBO J* 19(7): 1441-1449.

Clore, G. M. and C. D. Schwieters (2006). Concordance of residual dipolar couplings, backbone order parameters and crystallographic B-factors for a small alpha/beta protein: a unified picture of high probability, fast atomic motions in proteins. *J Mol Biol* 355(5): 879-886.

Coccetti, P., V. Zinzalla, G. Tedeschi, G. L. Russo, S. Fantinato, O. Marin, L. A. Pinna, M. Vanoni and L. Alberghina (2006). Sic1 is phosphorylated by CK2 on Ser201 in budding yeast cells. *Biochem Biophys Res Commun* 346(3): 786-793.

Csermely, P., T. Korcsmaros, H. J. Kiss, G. London and R. Nussinov (2013). Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol Ther* 138(3): 333-408.

Csermely, P., Korcsmáros, T., Kiss, H. J. M., London, G., and Nussinov, R. (2013). Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol. Ther.* 138, 333–408.

Csermely, P., K. S. Sandhu, E. Hazai, Z. Hoksza, H. J. Kiss, F. Miozzo, D. V. Veres, F. Piazza and R. Nussinov (2012). Disordered proteins and network disorder in network descriptions of protein structure, dynamics and function: hypotheses and a comprehensive review. *Curr Protein Pept Sci* 13(1): 19-33.

Damaschun, G., H. Damaschun, K. Gast and D. Zirwer (1999). Proteins can adopt totally different folded conformations. *J Mol Biol* 291(3): 715-725.

del Sol, A., C. J. Tsai, B. Ma and R. Nussinov (2009). The origin of allosteric functional modulation: multiple pre-existing pathways. *Structure* 17(8): 1042-1050.

Delmotte, A., E. W. Tate, S. N. Yaliraki and M. Barahona (2011). Protein multi-scale organization through graph partitioning and robustness analysis: application to the myosin-myosin light chain interaction. *Phys Biol* 8(5): 055010.

Dodson, G. G., D. P. Lane and C. S. Verma (2008). Molecular simulations of protein dynamics: new windows on mechanisms in biology. *EMBO Rep* 9(2): 144-150.

Doncheva, N. T., Y. Assenov, F. S. Domingues and M. Albrecht (2012). Topological analysis and interactive visualization of biological networks and protein structures. *Nat Protoc* 7(4): 670-685.

Doncheva, N. T., K. Klein, F. S. Domingues and M. Albrecht (2011). Analyzing and visualizing residue networks of protein structures. *Trends Biochem Sci* 36(4): 179-182.

Dror, R. O., R. M. Dirks, J. P. Grossman, H. Xu and D. E. Shaw (2012). Biomolecular simulation: a computational microscope for molecular biology. *Annu Rev Biophys* 41: 429-452.

Dyson, H. J. and P. E. Wright (2005). Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 6(3): 197-208.

Esteban-Martin, S., R. B. Fenwick and X. Salvatella (2010). Refinement of ensembles describing unstructured proteins using NMR residual dipolar couplings. *J Am Chem Soc* 132(13): 4626-4632.

Fenwick, R. B., S. Esteban-Martin and X. Salvatella (2011). Understanding biomolecular motion, recognition, and allostery by use of conformational ensembles. *Eur Biophys J* 40(12): 1339-1355.

Frauenfelder, H. and B. McMahon (1998). Dynamics and function of proteins: the search for general concepts. *Proc Natl Acad Sci U S A* 95(9): 4795-4797.

Frauenfelder, H., S. G. Sligar and P. G. Wolynes (1991). The energy landscapes and motions of proteins. *Science* 254(5038): 1598-1603.

Ghosh A., Vishveshwara S. (2007). A study of communication pathways in methionyl- tRNA synthetase by molecular dynamics simulations and structure network analysis. *Proc Natl Acad Sci U S A*. 104(40):15711-6.

Ghosh, A. and S. Vishveshwara (2008). Variations in clique and community patterns in protein structures during allosteric communication: investigation of dynamically equilibrated structures of methionyl tRNA synthetase complexes. *Biochemistry* 47(44): 11398-11407.

Goodey, N. M. and S. J. Benkovic (2008). Allosteric regulation and catalysis emerge via a common route. *Nat Chem Biol* 4(8): 474-482.

Greene, L. H. and V. A. Higman (2003). Uncovering network systems within protein structures. *J Mol Biol* 334(4): 781-791.

Harvey, M. J., Giupponi, G., and Fabritiis, G. De (2009). ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *J. Chem. Theory Comput.* 5, 1632–1639.

Hilser, V. J. (2010). Biochemistry. An ensemble view of allostery. *Science* 327(5966): 653-654.

Henzler-Wildman, K., and Kern, D. (2007). Dynamic personalities of proteins. *Nature* 450, 964–72.

Hodge, A. and M. Mendenhall (1999). The cyclin-dependent kinase inhibitory domain of the yeast Sic1 protein is contained within the C-terminal 70 amino acids. *Mol Gen Genet* 262(1): 55-64.

Hunenberger, P. H., A. E. Mark and W. F. van Gunsteren (1995). Fluctuation and cross-correlation analysis of protein motions observed in nanosecond molecular dynamics simulations. *J Mol Biol* 252(4): 492-503.

Invernizzi, G., M. Tiberti, M. Lambrughi, K. Lindorff-Larsen and E. Papaleo (2014). Communication routes in ARID domains between distal residues in helix 5 and the DNA-binding loops. *PLoS Comput Biol* 10(9): e1003744.

Isernia, C., E. Bucci, M. Leone, L. Zaccaro, P. Di Lello, G. Digilio, S. Esposito, M. Saviano, B. Di Blasio, C. Pedone, P. V. Pedone and R. Fattorusso (2003). NMR structure of the single QALGGH zinc finger domain from the Arabidopsis thaliana SUPERMAN protein. *Chembiochem* 4(2-3): 171-180.

Johnson, L. N. and R. J. Lewis (2001). Structural basis for control by phosphorylation. *Chem Rev* 101(8): 2209-2242.

Kannan, N. and S. Vishveshwara (1999). Identification of side-chain clusters in protein structures by a graph spectral method. *J Mol Biol* 292(2): 441-464.

Kar, G., Keskin, O., Gursoy, A., and Nussinov, R. (2010). Allostery and population shift in drug discovery. *Curr. Opin. Pharmacol.* 10, 715–22.

Karplus, M., Y. Q. Gao, J. Ma, A. van der Vaart and W. Yang (2005). Protein structural transitions and their functional role. *Philos Trans A Math Phys Eng Sci* 363(1827): 331-355; discussion 355-336.

Karplus, M. and J. Kuriyan (2005). Molecular dynamics and protein function. *Proc Natl Acad Sci U S A* 102(19): 6679-6685.

Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O., and Shaw, D. E. (2009). Long-timescale molecular dynamics simulations of protein structure and function. *Curr. Opin. Struct. Biol.* 19, 120–7.

Kitao, A. and N. Go (1999). Investigating protein dynamics in collective coordinate space. *Curr Opin Struct Biol* 9(2): 164-169.

Koivomagi, M., E. Valk, R. Venta, A. Iofik, M. Lepiku, E. R. Balog, S. M. Rubin, D. O. Morgan and M. Loog (2011). Cascades of multisite phosphorylation control Sic1 destruction at the onset of S phase. *Nature* 480(7375): 128-131.

Konrat, R. (2009). The protein meta-structure: a novel concept for chemical and molecular biology. *Cell Mol Life Sci* 66(22): 3625-3639.

Lambrughi, M., E. Papaleo, L. Testa, S. Brocca, L. De Gioia and R. Grandori (2012). Intramolecular interactions stabilizing compact conformations of the intrinsically disordered kinase-inhibitor domain of Sic1: a molecular dynamics investigation. *Front Physiol* 3: 435.

Lindorff-Larsen, K., R. B. Best, M. A. Depristo, C. M. Dobson and M. Vendruscolo (2005). Simultaneous determination of protein structure and dynamics. *Nature* 433(7022): 128-132.

Lindorff-Larsen, K., S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror and D. E. Shaw (2010). Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* 78(8): 1950-1958.

Lindorff-Larsen, K., S. Piana, R.O. Dror, D.E. Shaw (2011). How fast-folding proteins fold. *Science,* 334(6055):517-20.

Lindorff-Larsen, K., Maragakis, P., Piana, S., Eastwood, M. P., Dror, R. O., and Shaw, D. E. (2012). Systematic validation of protein force fields against experimental data. *PLoS One* 7, e32131.

Loh, S. N. (2010). The missing zinc: p53 misfolding and cancer. *Metallomics* 2(7): 442-449.

Ma, B. and R. Nussinov (2010). Enzyme dynamics point to stepwise conformational selection in catalysis. *Curr Opin Chem Biol* 14(5): 652-659.

Malgieri, G., L. Zaccaro, M. Leone, E. Bucci, S. Esposito, I. Baglivo, A. Del Gatto, L. Russo, R. Scandurra, P. V. Pedone, R. Fattorusso and C. Isernia (2011). Zinc to cadmium replacement in the A. thaliana SUPERMAN Cys(2) His(2) zinc finger induces structural rearrangements of typical DNA base determinant positions. *Biopolymers* 95(11): 801-810.

Manley, G., and Loria, J. P. (2012). NMR insights into protein allostery. *Arch. Biochem. Biophys.* 519, 223–31.

Mereghetti, P., L. Riccardi, B. O. Brandsdal, P. Fantucci, L. De Gioia and E. Papaleo (2010). Near native-state conformational landscape of psychrophilic and mesophilic enzymes: probing the folding funnel model. *J Phys Chem B* 114(22): 7609-7619.

Mitchell, E. M., P. J. Artymiuk, D. W. Rice and P. Willett (1990). Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J Mol Biol* 212(1): 151-166.

Mittag, T., S. Orlicky, W. Y. Choy, X. Tang, H. Lin, F. Sicheri, L. E. Kay, M. Tyers and J. D. Forman-Kay (2008). Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. *Proc Natl Acad Sci U S A* 105(46): 17772-17777.

Morra, G., G. Verkhivker and G. Colombo (2009). Modeling signal propagation mechanisms and ligand-based conformational dynamics of the Hsp90 molecular chaperone full-length dimer. *PLoS Comput Biol* 5(3): e1000323.

Motlagh, H. N., Li, J., Thompson, E. B., and Hilser, V. J. (2012). Interplay between allostery and intrinsic disorder in an ensemble. *Biochem. Soc. Trans.* 40, 975–80.

Nash, P., X. Tang, S. Orlicky, Q. Chen, F. B. Gertler, M. D. Mendenhall, F. Sicheri, T. Pawson and M. Tyers (2001). Multisite phosphorylation of a CDK inhibitor sets a threshold for the onset of DNA replication. *Nature* 414(6863): 514-521.

Nussinov, R., and Tsai, C.-J. (2013). Allostery in disease and in drug discovery. *Cell* 153, 293–305.

Palazzesi, F., Prakash, M. K., Bonomi, M., and Barducci, A. (2014). Accuracy of current all-atom force-fields in modeling protein disordered states. *J. Chem. Theory Comput., in press.*

Pandini, A., A. Fornili, F. Fraternali and J. Kleinjung (2013). GSATools: analysis of allosteric communication and functional local motions using a structural alphabet. *Bioinformatics* 29(16): 2053-2055.

Papaleo, E., K. Lindorff-Larsen and L. De Gioia (2012). Paths of long-range communication in the E2 enzymes of family 3: a molecular dynamics investigation. *Phys Chem Chem Phys* 14(36): 12515-12525.

Papaleo, E., P. Mereghetti, P. Fantucci, R. Grandori and L. De Gioia (2009). Free-energy landscape, principal component analysis, and structural clustering to identify representative conformations from molecular dynamics simulations: the myoglobin case. *J Mol Graph Model* 27(8): 889-899.

Papaleo, E., M. Pasi, L. Riccardi, I. Sambi, P. Fantucci and L. De Gioia (2008). Protein flexibility in psychrophilic and mesophilic trypsins. Evidence of evolutionary conservation of protein dynamics in trypsin-like serine-proteases. *FEBS Lett* 582(6): 1008-1018.

Papaleo, E., G. Renzetti and M. Tiberti (2012). Mechanisms of intramolecular communication in a hyperthermophilic acylaminoacyl peptidase: a molecular dynamics investigation. *PLoS One* 7(4): e35686.

Papaleo, E., Sutto L., Gervasio, F. L., and Lindorff-Larsen, K. (2014). Conformational Changes and Free Energies in a Proline Isomerase. *J. Chem. Theory Comput.*, 10: 4169-4174.

Pasi, M., L. Riccardi, P. Fantucci, L. De Gioia and E. Papaleo (2009). Dynamic properties of a psychrophilic alpha-amylase in comparison with a mesophilic homologue. *J Phys Chem B* 113(41): 13585-13595.

Pasi, M., M. Tiberti, A. Arrigoni and E. Papaleo (2012). xPyder: a PyMOL plugin to analyze coupled residues and their networks in protein structures. *J Chem Inf Model* 52(7): 1865-1874.

Piazza, F. and Y. H. Sanejouand (2008). Discrete breathers in protein structures. *Phys Biol* 5(2): 026001.

Piazza, F. and Y. H. Sanejouand (2009). Long-range energy transfer in proteins. *Phys Biol* 6(4): 046014.

Popovych, N., Sun, S., Ebright, R. H., and Kalodimos, C. G. (2006). Dynamically driven protein allostery. *Nat. Struct. Mol. Biol.* 13, 831–8.

Scarabelli, G., G. Morra and G. Colombo (2010). Predicting interaction sites from the energetics of isolated proteins: a new approach to epitope mapping. *Biophys J* 98(9): 1966-1975.

Seeber, M., A. Felline, F. Raimondi, S. Muff, R. Friedman, F. Rao, A. Caflisch and F. Fanelli (2011). Wordom: a user-friendly program for the analysis of molecular structures, trajectories, and free energy surfaces. *J Comput Chem* 32(6): 1183-1194.

Shaw, D. E., Grossman, J. P., Bank, J. A., Batson, B., Butts, J. A., Chao, J. C., Deneroff, M. M., Dror, R. O., Even, A., Fenton, C. H., et al. (2014). Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. IEE Press Piscataway, NJ, USA, 41–53.

Stacklies, W., F. Xia and F. Grater (2009). Dynamic allostery in the methionine repressor revealed by force distribution analysis. *PLoS Comput Biol* 5(11): e1000574.

Sugita, Y., and Y. Okamoto. 1999. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett*. 314:141–151.

Sutto, L. and C. Camilloni (2012). From A to B: a ride in the free energy surfaces of protein G domains suggests how new folds arise. *J Chem Phys* 136(18): 185101.

Szalay-Beko, M., R. Palotai, B. Szappanos, I. A. Kovacs, B. Papp and P. Csermely (2012). ModuLand plug-in for Cytoscape: determination of hierarchical layers of overlapping network modules and community centrality. *Bioinformatics* 28(16): 2202-2204.

Tehver, R., J. Chen and D. Thirumalai (2009). Allostery wiring diagrams in the transitions that drive the GroEL reaction cycle. *J Mol Biol* 387(2): 390-406.

Teilum, K., J. G. Olsen and B. B. Kragelund (2011). Protein stability, flexibility and function. *Biochim Biophys Acta* 1814(8): 969-976.

Tiberti, M., G. Invernizzi, M. Lambrughi, Y. Inbar, G. Schreiber and E. Papaleo (2014). PyInteraph: a framework for the analysis of interaction networks in structural ensembles of proteins. *J Chem Inf Model* 54(5): 1537-1551.

Tiberti, M. and E. Papaleo (2011). Dynamic properties of extremophilic subtilisin-like serine-proteases. *J Struct Biol* 174(1): 69-83.

Tompa, P. (2005). The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett* 579(15): 3346-3354.

Tsai, C.-J., and Nussinov, R. (2014). A unified view of "how allostery works". *PLoS Comput. Biol.* 10, e1003394.

Uversky, V. N. (2002). Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* 11(4): 739-756.

Uversky, V. N., J. R. Gillespie and A. L. Fink (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* 41(3): 415-427.

Varedi, K. S., A. C. Ventura, S. D. Merajver and X. N. Lin (2010). Multisite phosphorylation provides an effective and flexible mechanism for switch-like protein degradation. *PLoS One* 5(12): e14029.

Vendruscolo, M. (2007). Determination of conformationally heterogeneous states of proteins. *Curr Opin Struct Biol* 17(1): 15-20.

Vendruscolo, M., N. V. Dokholyan, E. Paci and M. Karplus (2002). Small-world view of the amino acids that play a key role in protein folding. *Phys Rev E Stat Nonlin Soft Matter Phys* 65(6 Pt 1): 061910.

Vendruscolo, M. (2007). Determination of conformationally heterogeneous states of proteins. *Curr. Opin. Struct. Biol.* 17, 15–20.

Verma, R., R. M. Feldman and R. J. Deshaies (1997). SIC1 is ubiquitinated in vitro by a pathway that requires CDC4, CDC34, and cyclin/CDK activities. *Mol Biol Cell* 8(8): 1427-1437.

Vishveshwara, S., Ghosh, A., and Hansia, P. (2009). Intra and inter-molecular communications through protein structure network. *Curr. Protein Pept. Sci.* 10, 146–60.

Whitley, M. J., and Lee, A. L. (2009). Frameworks for understanding long-range intra-protein communication. *Curr. Protein Pept. Sci.* 10, 116–27.

Xu, J., J. Reumers, J. R. Couceiro, F. De Smet, R. Gallardo, S. Rudyak, A. Cornelis, J. Rozenski, A. Zwolinska, J. C. Marine, D. Lambrechts, Y. A. Suh, F. Rousseau and J. Schymkowitz (2011). Gain of function of mutant p53 by coaggregation with multiple tumor suppressors. *Nat Chem Biol* 7(5): 285-295.

2 Studies of intrinsically disordered domains by molecular dynamics simulations integrated with experimental biophysical techniques.

2.1 Introduction

The tight relationship between structure and function has been a central dogma in protein biology over the last century. Recently, however, several examples emerged of proteins that are either entirely disordered or feature extended disordered regions. These intrinsically disordered proteins (IDPs) possess a number of crucial biological functions, yet do not conform to the prevailing structure-function view (Dunker et al. 2013). Due to their substantial differences from proteins characterized by a defined three-dimensional architecture, the emergence of IDPs has challenged the current methods in protein structural biology. IDPs are one of the more challenging cases of study for both computational and experimental structural biology. IDPs are very common in nature and the proportion of proteins containing disordered regions increases with the complexity of an organism (Ward et al., 2004) especially in eukaryotes, where ~50% of the proteins are predicted to contain long disordered regions, and ~30% are classified as IDPs (Uversky and Dunker, 2010). Further, more than one third of the human proteins contain sequences longer than 30 aminoacids predicted to be disordered. IDPs are essential in cell, since they carry out key roles in a large number of biological processes in crucial areas such as transcriptional regulation, translation and cellular signal transduction, protein phosphorylation, storage, assembly of large multiprotein complexes and recently as chaperone (Uversky et al., 2000, Uversky, 2002, Dyson and Wright, 2005, Uversky and Dunker, 2010, Tompa and Kovacs, 2010). Further, impairment of their function is often associated with severe diseases (Dyson and Wright, 2005). It has been suggested that proteins involved in eukaryotic signal transduction and associated with cancer have an increased propensity for intrinsic disorder (Iakoucheva et al., 2002). IDPs lack a well-defined and organized 3D structure in their free state and instead consist of a wide and fluctuating ensemble of conformations. IDPs usually have low hydrophobic bulky aminoacids, which normally form the core of folded globular proteins, and a high number of polar and charged residues (such as Gln, Ser, Pro, Glu, Lys) (Vucetic, et al., 2003). In general, intrinsically disordered sequences cannot bury sufficient hydrophobic core to fold into a stable structures but IDPs, also in their unbound state, can populate collapsed and disordered molten globule-like conformations, stabilized by intramolecular interactions such as electrostatic and hydrophobic interactions (Espinoza-Fonseca, 2009, Wostenberg et al., 2011). Nevertheless, the existence of highly collapsed states for IDPs in solutions is largely debated. Moreover, most IDPs can transiently populate partially structured conformations, also in their unbound state, and display intrinsically folded structural units (IFSUs), regions of the protein that have higher propensity to adopt isolated and fluctuating elements of secondary structure. These elements are often helical and are thought to provide seeds of binding interfaces (Sivakolundu et al., 2005, Espinoza-Fonseca et al., 2007, Wright and Dyson, 2009, Kjaergaard et al., 2010, Norholm et al., 2011). Moreover, it has been proposed that IDPs rarely behave as true random coils, especially in non-denaturing media, since even in highly unfolded states they still show propensity to form local elements of secondary structure or small hydrophobic clusters (Dyson and Wright, 2004). Structural properties of IDPs observed in solution may be also important to determine theirbiological function *in vivo*. Indeed, IDPs can exert their physiological roles in a multi-functional way thank to their structural heterogeneity which permits to make interactions with several different binding partners, but these mechanisms are still to understand. Generally IDPs and disordered domains undergo transitions to more ordered states or fold into stable secondary or tertiary structures upon interactions with binding partners, called coupled folding and binding processes (Wright and Dyson 1999, Dyson and Wright, 2005, Espinoza-Fonseca, 2009) although cases are known in which structural disorder is retained also in the bound state (Tompa and Fuxreiter, 2008). Despite disorder to order transitions of IDPs have a high entropic cost, the reaction is thermodynamic driven by favorable enthalpic contribution from the binding event. In this process IDPs form complexes with high specificity and relatively low affinity, which is essential for proteins involved in signal transduction. Indeed, these proteins have to bind with high specificity to a target but also have to fast dissociate to modulate the initiation and termination of the signaling process. Moreover, coupled folding and binding processes permit to maintain conformational flexibility that facilitates the regulation by post-translational modification, through the binding with both physiological target and modifying enzymes. Indeed the binding of IDPs to their targets and their biological functions are often regulated by covalent post-translational modifications. The relative instability of the IDPs provides a further level of cellular control through proteolytic degradation, permitting an additional regulatory level over their functions. Disordered regions can also have a biological cost since they are site susceptible to mutations and associated with the promotion of protein folding diseases and several neurodegenerative diseases, for example prion, Parkinson’s and polyglutammine (polyQ) disease are linked with disordered proteins.

Considerable effort has been devoted to characterize IDP structural and dynamic at the atomic level (Dyson and Wright, 2005; Dunker et al., 2008, Turoverov et al., 2010; Fisher and Stultz, 2011) but lot is still unknown. Moreover, to understand the importance of disorder in biological functions of IDPs, such as molecular recognition, it is required not only the description of the bound states (Morin et al., 2006; Receveur-Bréchot et al., 2006; Espinoza-Fonseca, 2009b; Hazy and Tompa, 2009; Wright and Dyson, 2009), but also the characterization of the elusive and heterogeneous unbound states (Dunker et al., 2008; Salmon et al., 2010; Fisher and Stultz, 2011; Szasz et al., 2011; Bernado and Svergun, 2012; Schneider et al., 2012). Different biophysical techniques have been successfully exploited to thoroughly characterize IDPs in solution but the highly dynamical behavior of IDPs makes them difficult to study with only *in vitro* experiments. In this context, computational approaches and atomistic molecular dynamics (MD) simulations proved suitable to describe the conformational landscape of IDPs or denatured proteins, identifying intrinsic/residual structure, binding and folding processes (Espinoza-Fonseca, 2009, Rauscher and Pomes, 2010; Cino et al., 2011, Fisher and Stultz, 2011, Arrigoni et al., 2012, Ganguly et al., 2012; Lindorff-Larsen et al., 2012, Knott and Best, 2014), even if limitations are still present in MD simulations especially due to risk of overcompaction of IDPs during the sampling. Thus, the comparison and validation with experiments in necessary. In this thesis, we present a multidisciplinary approach, based on the integration of canonical MD simulations with biophysical techniques, such as Electro-Spray Ionization Mass Spectrometry (ESI-MS) and Nuclear magnetic Resonance (NMR) to characterize the free ensembles of ID domains in solution in atomistic details Atomistic MD simulations have been analysed by several tools to identify structural properties of the IDP ensemble, ranging from Principal Component Analysis, and methods inspired by graph theory. Our investigations focused on the C-terminal domain of yeast Sic1 (Brocca et al., 2012) (Chap.2.2) and the disordered regions of human Ataxin-3 (Saunders et al., 2009): region from residue 182 to 291 (Chap2.3 and 2.4) C-terminal region from residue 306 to 360 (Chap. 2.5), to provide a description of their heterogeneity in solution and identify structures that resemble the conformations bound to their biological partners.

References

Arrigoni, A., Grillo, B., Vitriolo, A., De Gioia, L., and Papaleo, E. (2012). C-terminal acidic domain of ubiquitin-conjugating enzymes: a multi-functional conserved intrinsically disordered domain in family 3 of E2 enzymes. *J. Struct. Biol.* 178, 245–259.

Brocca, L., J. Cannavino, L. Coletto, G. Biolo, M. Sandri, R. Bottinelli and M. A. Pellegrino (2012). The time course of the adaptations of human muscle proteome to bed rest and the underlying mechanisms. *J Physiol* 590(Pt 20): 5211-5230.

Cino, E. A., J. Wong-ekkabut, M. Karttunen and W. Y. Choy (2011). Microsecond molecular dynamics simulations of intrinsically disordered proteins involved in the oxidative stress response. *PLoS One* 6(11): e27371.

Dunker, A. K. and Z. Obradovic (2001). The protein trinity--linking function and disorder. *Nat Biotechnol* 19(9): 805-806.

Dunker A.K., M. Madan Babu, E. Barbar, M. Blackledge, S.E. Bondos, Z. Dosztányi, H.J. Dyson, J. Forman-Kay, M. Fuxreiter, J. Gsponer, K.-H. Han, D.T. Jones, S. Longhi, S.J.Metallo, K. Nishikawa, R. Nussinov, Z. Obradovic, R.V. Pappu, B. Rost, P. Selenko, V. Subramaniam, J.L. Sussman, P. Tompa, V.N. Uversky, (2013), *Intrinsically Disordered Proteins* 1:1, e24157.

Dyson, H. J. and P. E. Wright (2002). Coupling of folding and binding for unstructured proteins. *Curr Opin Struct Biol* 12(1): 54-60.

Dyson, H. J. and P. E. Wright (2004). Unfolded proteins and protein folding studied by NMR. *Chem Rev* 104(8): 3607-3622.

Dyson, H. J. and P. E. Wright (2005). Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 6(3): 197-208.

Espinoza-Fonseca, L. M. (2009). Leucine-rich hydrophobic clusters promote folding of the N-terminus of the intrinsically disordered transactivation domain of p53. *FEBS Lett* 583(3): 556-560.

Espinoza-Fonseca, L. M. (2009). Reconciling binding mechanisms of intrinsically disordered proteins. *Biochem Biophys Res Commun* 382(3): 479-482.

Espinoza-Fonseca, L. M. (2009). Thermodynamic aspects of coupled binding and folding of an intrinsically disordered protein: a computational alanine scanning study. *Biochemistry* 48(48): 11332-11334.

Espinoza-Fonseca, L. M., D. Kast and D. D. Thomas (2007). Molecular dynamics simulations reveal a disorder-to-order transition on phosphorylation of smooth muscle myosin. *Biophys J* 93(6): 2083-2090.

Ganguly, D., Zhang, W., and Chen, J. (2012). Synergistic folding of two intrinsically disordered proteins: searching for conformational selection. *Mol. Biosyst*. 8, 198–209.

Hazy, E. and P. Tompa (2009). Limitations of induced folding in molecular recognition by intrinsically disordered proteins. *Chemphyschem* 10(9-10): 1415-1419.

Iakoucheva, L. M., C. J. Brown, J. D. Lawson, Z. Obradovic and A. K. Dunker (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* 323(3): 573-584.

Knott, M., Best R.B. (2014) Discriminating binding mechanisms of an intrinsically disordered protein via a multi-state coarse-grained model. *J Chem Phys* 140(17):175102

Kjaergaard, M, Teilum, K, Poulsen, F.M. (2010) Conformational selection in the molten globule state of the nuclear coactivator binding domain of CBP. *Proc Natl Acad Sci U S A*. 107(28):12535-40.

Lindorff-Larsen, K., Trbovic, N., Maragakis, P., Piana, S., and Shaw, D. E. (2012). Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *J. Am. Chem. Soc.* 134, 3787–3791.

Norholm A.-B., Hendus-Altenburger R., Bjerre G., Kjaergaard M., Pedersen S. F., Kragelund B. B. (2011). The intracellular distal tail of the Na(+)/H(+) exchanger NHE1 Is intrinsically disordered: implications for NHE1 trafficking. *Biochemistry* 50, 3469–3480

Rauscher, S. and R. Pomes (2010). Molecular simulations of protein disorder. *Biochem Cell Biol* 88(2): 269-290.

Sivakolundu, S. G., D. Bashford and R. W. Kriwacki (2005). Disordered p27Kip1 exhibits intrinsic structure resembling the Cdk2/cyclin A-bound conformation. *J Mol Biol* 353(5): 1118-1128.

Tompa P, Kovacs D. (2010). Intrinsically disordered chaperones in plants and animals. *Biochem Cell Biol*. 88(2):167-74.

Tompa, P. and M. Fuxreiter (2008). Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem Sci* 33(1): 2-8.

Turoverov, K. K., I. M. Kuznetsova and V. N. Uversky (2010). The protein kingdom extended: ordered and intrinsically disordered proteins, their folding, supramolecular complex formation, and aggregation. *Prog Biophys Mol Biol* 102(2-3): 73-84.

Uversky, V.N., J.R. Gillespie, A.L. Fink, (2000), Why are "natively unfolded" proteins unstructured under physiologic conditions?, *Proteins*, 41(3):415-27.

Uversky, V.N., (2002) Natively unfolded proteins: A point where biology waits for physics. *Protein Sci.*, 11(4): 739–756.

Uversky, V. N. and A. K. Dunker (2010). Understanding protein non-folding. *Biochim Biophys Acta* 1804(6): 1231-1264.

Vucetic, S., C. J. Brown, A. K. Dunker and Z. Obradovic (2003). Flavors of protein disorder. *Proteins* 52(4): 573-584.

Ward, J. J., J. S. Sodhi, L. J. McGuffin, B. F. Buxton and D. T. Jones (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 337(3): 635-645.

Wostenberg, C., S. Kumar, W. G. Noid and S. A. Showalter (2011). Atomistic simulations reveal structural disorder in the RAP74-FCP1 complex. *J Phys Chem B* 115(46): 13731-13739.

Wright, P. E. and H. J. Dyson (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* 293(2): 321-331.

Wright, P. E. and H. J. Dyson (2009). Linking folding and binding. *Curr Opin Struct Biol* 19(1): 31-38.

2.2 Intramolecular interactions stabilizing compact conformations of the intrinsically disordered kinase-inhibitor domain of Sic1: a molecular dynamics investigation.

Lambrughi M., Papaleo E., Testa L., Brocca S., De Gioia L. and Grandori R. (2012)
Front. Physiol. **3**:435.

2.2.1 Introduction

Cyclin-dependent kinase inhibitors (CKIs) are key regulatory proteins of the eukaryotic cell cycle, which modulate cyclin-dependent kinase (Cdk) activity (Besson et al., 2008; Barberis, 2012). CKIs perform their inhibitory effect by the formation of ternary complexes with a target kinase and its cognate cyclin. These regulators generally belong to the class of intrinsically disordered proteins (IDPs), which lack a well-defined and organized three-dimensional structure in their free state, undergoing folding upon binding to specific partners. Several IDPs are involved in cell cycle regulation (Galea et al., 2008). The IDP Sic1 is a yeast CKI that regulates the timing of entrance into S-phase by inhibition of the Clb5-6/Cdk1 complex (Schwob et al., 1994) preventing phosphorylation of G1 substrates and, therefore, entrance to the S phase. The G1-to-S transition and cell cycle progression is executed only upon ubiquitin-dependent Sic1 degradation by the proteasome. Sic1 is targeted for destruction by the Skp1-Cul1-F-box complex, SCF^{cdc4}. The interaction with SCF is triggered by Sic1 multiple phosphorylation in its N-terminal region (Nash et al., 2001; Mittag et al., 2008, Koivomagi et al., 2011) and inhibited by the stress-response phosphorylation on T173 (Escote et al., 2004; Yaakov et al., 2009). Such a network of regulatory functions contributes to effective cell division and genome integrity (Verma et al., 1997; Mendenhall and Hodge, 1998; Deshaies and Ferrell, 2001; Nash et al., 2001; Barberis, 2012; Barberis et al., 2012) (Figure 1).

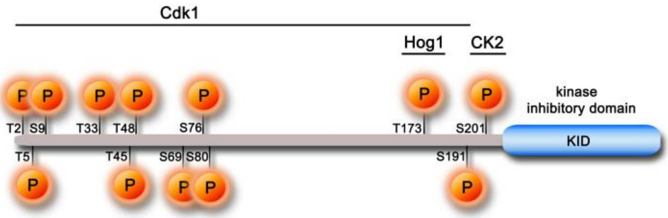


Figure 1. Schematic representation of the structure of the full-length Sic1 protein and its principal phosphorylation sites. The minimal functional KID fragment, corresponding to the last 70 residues of the protein, is indicated as a blue box. The main kinases that regulate Sic1 activity in the cell cycle are indicated, along with their phosphorylation sites on the whole protein.

Sic1 is also involved in other regulatory processes, such as exit from mitosis (Lopez-Aviles et al., 2009) and coupling of cell growth to cell-cycle progression (Coccetti et al., 2004; Barberis et al., 2007). Sic1 possess a wide range of regulatory function in the yeast cell cycle, both inhibitory and activatory, acting as a timer to control the oscillatory waves of Clb cyclins and regulate activities of the Clb/Cdk1 complexes, in order to ensure a precise temporal progression, guaranteeing an effective cell division and the integrity of genome (Barberis, 2012;Barberis et al., 2012). Sic1 is 284-residue long and, although disordered in its whole length, it transiently populates conformations with various degrees of compactness, revealing global structural properties more typical of a collapsed chain than of a random coil (Brocca et al., 2009; Brocca et al., 2011b). Its free state is characterized by dynamic tertiary and secondary structure (Brocca et al., 2009). The little content in secondary, mainly helical, structure is distributed quite uniformly throughout the polypeptide chain, although the C-terminus is slightly more ordered than the N-terminus (Brocca et al., 2011a; Brocca et al., 2011b). Conformational analysis by electrospray-ionization mass spectrometry (ESI-MS) and limited proteolysis suggests that the C-terminal moiety also contains more tertiary structure than its complementary N-terminal region (Brocca et al., 2011b; Testa et al., 2011b). The last 70 residues have been identified as the minimal protein fragment for in-vivo Cdk1 inhibition (Verma et al., 1997; Hodge and Mendenhall, 1999; Nash et al., 2001), and have been proposed to be structurally and functionally related to the kinase inhibitory domain (KID) of the mammalian tumor-suppressor p21 and p27 Kip/Cip proteins (Barberis et al., 2005; Toyoshima and Hunter, 1994) (Figure 1). An X-ray structure of the p27 ternary complex with Cdk2-cyclin A is available (Russo et al., 1996). No structural data are available for the yeast ternary complex, but a model was developed based on the template of the mammalian complex (Barberis et al., 2005). Sic1 KID has been identified as the minimal protein fragment necessary and sufficient for binding and inhibition of cyclin-CDK complexes in yeast cells (Verma et al., 1997;Hodge and Mendenhall, 1999;Nash et al., 2001) but the overall regulatory activity of Sic1 is highly controlled, also by multiple phosphorylations (Nash et al., 2001;Coccetti et al., 2006). IDPs are often been found to regulate key cellular processes and pathways of signal transduction through binding to multiple targets, due to the lack of a precise structures and to their wide conformational ensemble accessible and flexible nature (Uversky et al., 2000;Uversky, 2002;Dyson and Wright, 2005;Tompa, 2005). Also a wide range regulatory role can be proposed for Sic1, since multiple functional roles in cell and several interactions with different partners have been identified. Further, the central importance of IDPs in several crucial pathways entails to highly regulate the action of this protein, through modulation of the dynamical properties and the interactions with biomolecules target. For Sic1 G1-cyclins (Clns1-3)-Cdk1(Cdc28) complexes control the inhibitory activity of Sic1 on the Clb-Cdc28 kinases, through phosphorylation of multiple sites, localized at the N-terminal region, triggering Sic1 ubiquitination and degradation, mediated by the cdc34-SCF ubiquitin ligase complex (Verma et al., 1997) and a efficient and high affinity recognition minimally requires six of the nine phosphorylation sites localized in the N-terminal region of Sic1 (Nash et al., 2001). The cascades of multiple phosphorylation determine a effective mechanism for protein degradation, which seems to be common to several other proteins involved in signaling processes, that permits the cell to provide an efficient and fine regulations of cellular processes like cell cycle promotion, acting as a timing device for G1-phase progression and ensuring a precise control on S-phase and DNA replication start, also determining a effectiveness threshold level of Cln1, 2, 3/Cdc28 in late G1-phase and Clb/Cdc28 and other kinases (Varedi et al., 2010).In this context we investigated the conformational ensemble of Sic1 KID in its free and compact state, by carry out multiple all-atom molecular dynamics (MD) simulations in explicit solvent starting from extended models of Sic1 KID, collecting overall more than 500 ns of MD trajectories, and integrated by spectroscopic and spectrometric data collected by Dr. Lorenzo Testa in the group of Prof. Rita Grandori. The results describe intrinsic secondary and tertiary structure and helical IFSUs are identified, along with networks of intramolecular interactions. The results identify a group of hub residues and electrostatic interactions involved in the stabilization of globular states and suggest a predominant role of electrostatic interactions in promoting protein compaction. We have investigated the conformational ensemble of the isolated Sic1 KID fragment, providing the first atomic-level description of the compact states and dynamical behavior, in relationship with its biological functions.

2.2.2 Materials and Methods

Starting structures for MD simulations

An extended conformation of the Sic1 KID fragment (residues 215-284) was generated by the *generated_extended.inp* module of the *crystallography & NMR system* (CNS) software (Brunger, 2007) to avoid tertiary contacts and to allow the protein to rearrange during MD simulations. Starting from this structure, 10 different models (featuring a mainchain rmsd in the range from 0.7 to 10 nm when compared) were generated by the program *MODELLER*, including constraints derived from secondary-structure prediction. Both approaches were previously applied to other IDPs (Gardebien et al., 2006, Cino et al., 2011). In particular, the regions identified by at least two predictors (Barberis et al., 2005) have been constrained to α-helices (T226-L238 and I244-I248). *MODELLER* generates structural models satisfying spatial restraints. It employs knowledge-based probability density functions (PDFs) derived by statistical mechanics (Eswar et al., 2008). Among the models featuring no side-chain interactions, four models were selected as the starting structures of MD simulations based on the DOPE and GA341 *MODELLER* scores.

Molecular-dynamics simulations

MD simulations were performed using the 4.5.3 version of the *GROMACS* software (www.gromacs.org), implemented on a parallel architecture, using the GROMOS96 force field. The Sic1 KID models described above were used as starting structures for all-atom, explicit-solvent, MD simulations, employing periodic boundary conditions. The Sic1 KID molecule was soaked in a dodecahedral box of Simple Point Charge (SPC) water molecules (Fuhrmans et al., 2010) with all the protein atoms at a distance equal or greater than 1.5 nm from the box edges. Productive MD simulations were performed in the isothermal-isobaric (NPT) ensemble at 300K and 1 bar, using an external bath with thermal and pressure coupling of 0.1 and 1 ps, respectively. The LINCS algorithm (Hess et al., 1997) was used to constrain heavy-atom bonds, allowing for a 2 fs time-step. Electrostatic interactions were modeled by the Particle-mesh Ewald summation scheme (Darden et al., 1993). Van der Waals and Coulomb interactions were truncated at 1.2 nm, a cutoff value previously used for IDP simulations and experimentally validated by comparison with electronic paramagnetic resonance and fluorescence data (Espinoza-Fonseca et al., 2008, Espinoza-Fonseca, 2009a). The non-bonded interaction list was updated every 10 steps and conformations were stored every 2 ps. The simulations were carried out in the presence of Na^+ and Cl^- counterions, to simulate a physiological ionic strength (150 mM), according to a protocol previously employed for other IDPs (Espinoza-Fonseca, 2009a; Arrigoni et al., 2012). The length of each simulation (replicate, r.) ranged from 50 (r.1 - r.6 and r.9) to 100 ns (r.7 and r.8). The first 0.5 ns of each replicate were discarded to avoid artifacts arising from the preparation procedure. A concatenated macro-trajectory, including 9 replicates, for a total duration of 545.5 ns, was generated to obtain a conformational ensemble of Sic1 KID in solution.

Analysis of MD simulations

The secondary-structure content of the models was calculated by the *DSSP* program (Kabsch and Sander, 1983), along with a residue-dependent profile of secondary structure persistence. Salt-bridge, aromatic and hydrophobic interactions were analysed as previously described (Tiberti and Papaleo, 2011; Arrigoni et al., 2012). In particular, a persistence cutoff of 20% and a distance cutoff of 0.5 nm were employed. Aromatic interactions were analysed using a 0.6 nm distance cutoff.

Analysis of salt bridge interaction networks

Salt-bridge interactions were also analysed by the *Pymol* plugin *xPyder*, representing pairwise relationships related to protein structures as two-dimensional matrices (Pasi et al., 2012). In particular, the module for network analysis implemented in *xPyder* has been employed. A network is described as a set of points (nodes) and connections between them (edges), according to (Vishveshwara et al., 2009). A path is defined as a sequence of nodes for which an edge always exists between two consecutive nodes of the path. A matrix describing the persistence of each salt bridge has been used as input file. The program represents each residue of the matrix as a node of a simple, weighted graph connected by edges, whose weights are defined by the persistence of the interaction in the MD ensemble (i.e. the number of trajectory frames in which the interaction was present, over the total number of frames). Residues connected by more than 3 edges to their neighbors (Brinda and Vishveshwara, 2005) are referred to as hubs of the interaction networks. Hubs in protein networks are known to play key roles in protein structure or function (Vishveshwara et al., 2009, Angelova et al., 2011). The connected components of the graph were also calculated. These are isolated sub-graphs in which all the edges are linked by at least one path, but no path exists between the nodes of the connected component and the rest of the graph. This analysis allows us to identify different clusters of salt-bridge networks. Finally, all the possible paths existing between two nodes of the graph were calculated employing a variant of the depth-first search algorithm (Cormen et al., 2009), as implemented in *xPyder*. The searching procedure was carried out so that the same node is not visited more than once to avoid entrapment in cycles.

Principal component analysis (PCA) and Free Energy Landscape (FEL)

PCA highlights high-amplitude, concerted motions in MD trajectories, through the eigenvectors of the mass-weighted covariance matrix (C) of the atomic positional fluctuations (Amadei et al., 1993). Both all atom and Cα Cs were calculated for the macro-trajectory. Given a reaction coordinate q_α, the probability of finding the system in a particular state q_α is proportional to (e-$^{G(q\alpha)/kT}$), where $G(q_\alpha)$ is the Gibbs free energy of that state. The FEL can be computed from the equation $G(q_\alpha)=-kTln[P(q_\alpha)]$, where k is the Boltzmann constant, T is the temperature of the simulation and $P(q_\alpha)$ is an estimation of the probability density function obtained from a histogram of the MD data. Considering two different reaction coordinates, for example q and p, the two-dimensional FEL can be obtained from the joint probability distribution $P(q,p)$ of the considered variables. In particular, the reaction coordinates considered in this study were the first and the second, as well as the first and the third cartesian principal components (PCs or eigenvectors) derived by the PCA procedure described above. The evaluation of the conformational sampling convergence was performed calculating the root mean square inner product (rmsip) as a measure of similarity between the essential subspaces, obtained from the different independent simulations, defined by their basis vectors (Amadei et al., 1999):

$$\mathrm{RMSIP} = \frac{1}{D}\sum_{i=1}^{D}\sum_{i=1}^{D}(\eta_i^A \eta_j^B)$$

Where η_i^A and η_j^B are the eigenvectors to be compared and D the number of eigenvectors considered. The rmsip was computed on the first 10 eigenvectors (Amadei et al., 1999).

Cluster analysis

The structures belonging to each main basin on the FEL were isolated and the mainchain root mean square deviation (rmsd) matrix of each basin was calculated. The Gromos algorithm (Keller et al., 2010) was employed for clustering, using a cutoff of 0.4 nm. For each cluster, the structure with the lowest rmsd with respect to the other members of the cluster was selected as the average structure.

Order parameter O

The order parameter *O* was calculated according to the formula (Fisher and Stultz 2011):

$$O = \sum_{i=1}^{n} w_i \log_2 \left[1 + \sum_{j=1}^{n} w_j \exp\left(-\frac{D^2(s_i, s_j)}{2\langle D^2 \rangle} \right) \right]$$

where $D^2(s_i, s_j)$ is the Cα mean-square distance (MSD) between the structures s_i and s_j, and $<D^2>$ is the average pairwise MSD due to the fluctuations of a typical protein structure at a specific temperature (according to Fisher and Stultz, 2011, this value is 0.27 nm). Protein conformations were extracted from the MD macro-trajectory after clustering, by the Gromos algorithm using three different values of rmsd cutoff (0.3, 0.4 and 0.5 nm). The average structure of each cluster was extracted to create an ensemble of protein conformations and to calculate the order parameter *O*. The weights associated with each conformation were set as the relative size of each cluster, dividing the number of structures assigned to a cluster by the total number of frames in the trajectory. The results obtained by using the different cutoffs are reported in Table 2.

Mass spectrometry

Electrospray ionization mass spectrometry (ESI-MS) experiments were performed on a hybrid quadrupole-time-of-flight mass spectrometer (QSTAR ELITE, Applied Biosystems, Foster City, CA) equipped with a nano-ESI sample source. Samples of 10 μM recombinant Sic1 KID fused to a C–

terminal His_6 tag (Brocca et al., 2011b) in 50 mM ammonium acetate, pH 6.5 were loaded in metal-coated borosilicate capillaries for nanospray (Proxeon, Odense, DK) with medium-length emitter tip and 1 μm internal diameter. The instrument was calibrated using renine inhibitor (1757.9 Da) (Applied Biosystems) and its fragment (109.07 Da) as standards. Spectra were acquired in the 500–2000 *m/z* range, with accumulation time 1 s, ion-spray voltage 1200–1400 V, declustering potential 60–80 V, keeping the instrument interface at room temperature. Spectra were averaged over a time period of 2 min. Data analysis was performed by the program *Analyst QS 2.0* (Applied Biosystems). Gaussian fitting of ESI-MS spectra was carried out on row data reporting ion relative intensity *versus* charge (Dobo and Kaltashov, 2001). These data were fitted by the minimal number of Gaussian functions leading to a stable fit. Fitting analyses were performed by the software *OriginPro 7.5* (Originlab, Northampton, MA).

2.2.3 Results.

Structural models of Sic1 KID (residues 215-284) in an extended conformation were generated as described in the Materials and Methods, satisfying secondary-structure constraints according to the previously published prediction (Barberis et al., 2005). This prediction is also in agreement with the average content of secondary structure indicated by circular dichroism (CD) (Brocca et al., 2011a) and Fourier-transform infrared (FT-IR) spectroscopy (Brocca et al., 2011b). Multiple, 50-100 ns, independent MD simulations were carried out starting from four extended models, collecting overall more than 500 ns of MD trajectories (Figure 2). The results of MD simulations were analysed with reference to secondary-structure content, solvent-accessible surface (SAS), and intramolecular interactions as described in details in the following. Moreover, distinct conformational substates were identified in the simulated ensemble by integrating structural clustering, PCA and FEL calculations (Papaleo et al., 2009).

Order parameter O

The order parameter *O* was calculated to evaluate the heterogeneity of the conformational ensemble described by the MD simulations, (Fisher and Schultz, 2011). Indeed, this parameter can be considered as a quantitative measure of the disorder in a given structural ensemble. The *O* parameter was calculated on the average structures derived from cluster analysis on the MD ensemble. The results are reported in Table 1. The Sic1 KID fragment displays very low values of the *O* parameter, ranging from 0.141 to 0.156, depending on the number of clusters considered. The limit value of 0 applies to the ideal case of an infinite number of equally populated, different conformations. Therefore, these results indicate that Sic1 KID exists as a highly heterogeneous conformational ensemble, confirming the strong propensity of this fragment for structural disorder.

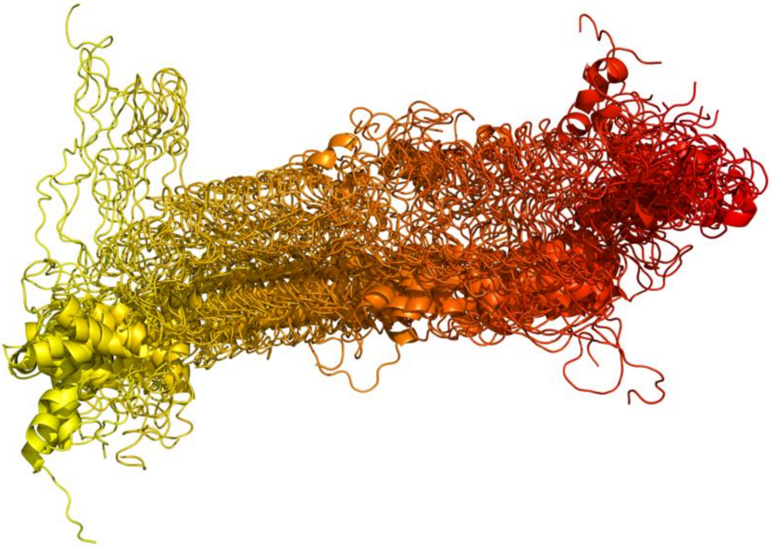


Figure 2. Structural displacement of the Sic1 KID domain in the MD macro-trajectory of more than 0.5 μs. The equilibrated parts of each replicate were concatenated in a single macro-trajectory, shown here. Snapshots are collected every 2 ns, overlaid and represented by a color gradient, from yellow (the starting structure of replicate 1) to red (the last frame of replicate 9).

Table 1. Order parameter estimated for the MD ensemble of Sic1 KID using three different cutoff values.

Cutoff of clustering	*Number of clusters*	*Order parameter*
6	55	0.141
5	79	0.144
4	158	0.156

Dynamic behavior

To evaluate the conformational sampling achieved by our MD investigation, and to better define the dynamic behavior of Sic1 KID, all-atom and Cα PCA were carried out on the macro-trajectory. PCA can provide an estimate of the conformational sampling achieved in a MD ensemble (Hess, 2002) and it can offer a representation of the sampled conformational landscape. The projections of the MD macro-trajectory on the first two eigenvectors show a good sampling of the conformational space. In fact, the different simulations sampled both different regions of the essential subspace, as well as they feature a partial overlap (Figure 3A). In particular, simulations starting from different initial models often populate the same basins, confirming that the MD ensemble here described provide a good description of the structural landscape (Amadei et al., 1999; Hess, 2002; Papaleo et al., 2009). Similar results were also achieved analyzing the two-dimensional projections along the first and the third principal components, as well as the second and the third PCs (Figure 3B-C).

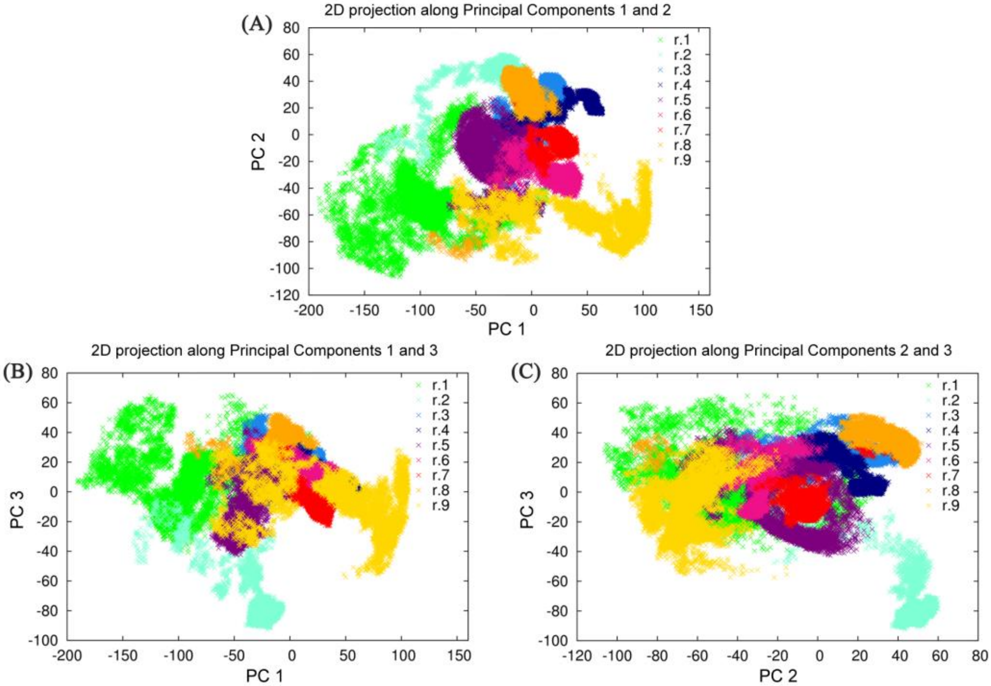


Figure 3. Projections of the MD trajectories of Sic1 KID on the first 3 eigenvectors derived from PCA of the macro-trajectory. A) First (x axis) and second (y axis) eigenvectors, B) first (x axis) and third (y axis) eigenvectors, C) second (x axis) and third (y axis) eigenvectors. The distinct contributions of the MD replicates are shown by different colors.

To better quantify the overlap, the rmsip of the first 10 PCs between the different replicates (Figure 4A) was calculated as described in Materials and Methods. It is an index of similarity between essential subspaces and it ranges from 0.33 to 0.55 in our simulations, with an average value of 0.44.

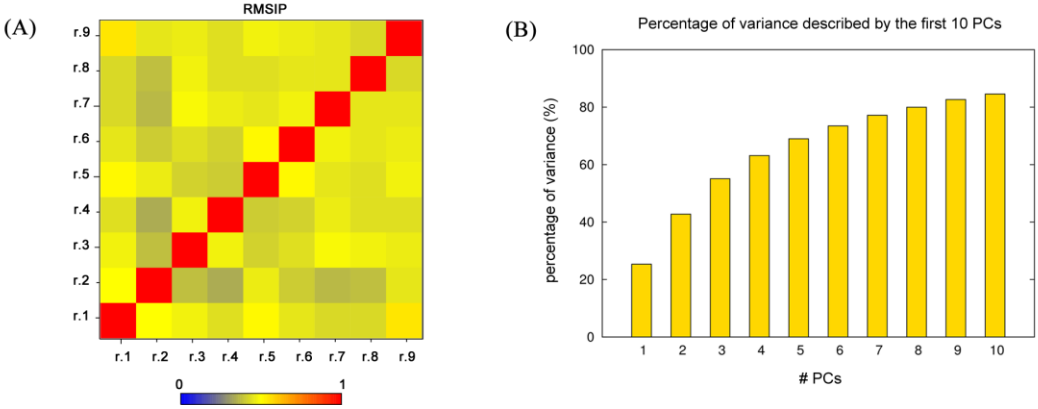


Figure 4. PCA analysis. A) Rmsip matrix calculated for each pairwise comparisons of the single replicate, using the conformational space described by the first 10 principal components. The Rmsip values are highlighted by a color gradient, from blue to red. B) Percentage of the cumulative variance described by the first ten eigenvectors. The first three eigenvectors account together for more than 50% of the global motions of the domain, while the first five describe more than 70% of them.

Overall, rmsip indicates a satisfactory convergence of the simulations with a high overlap of the dynamics information provided by each of them. Since the first three PCs account alone for more than 50% of the global motion of Sic1 KID (Figure 4B), they can be used as the so-called *essential subspace* to describe the main dynamical properties of this domain. The projections on the first three eigenvectors show that the protein has a very heterogeneous structural ensemble composed by several different, highly populated conformations (Figure 3).

To characterize and isolate the different conformational states of the MD ensemble, the FEL was calculated (Zhuravlev et al., 2009) using the first two PCs principal components as reaction coordinates (Figure 5). An average structure representative of each FEL basin is also reported in Figure 5. It is the structure, which feature the lowest mainchain rmsd with respect to all the other structures in the same basin, as detailed in Materials and Methods. Other FEL representations were also calculated for comparison, using the first and the third, or the second and the third PCs (*data not shown*). It should be pointed out that these low-dimensional FEL representations, even if not sufficiently accurate to calculate the free-energy barriers between conformational basins and to quantify the transitions between the different substates, are very useful to describe the conformational landscape accessible to the molecule in the MD ensemble. In the FEL representation depicted in Figure 5, seven major (A, B, C, D, E, F, I) and three minor (G, H, L) conformational basins can be identified (Figure 5). The projections of the single trajectories on the FEL (Figure 6) show that each replicate can sample different conformational basins, as was also evident by the analyses in Figure 3.

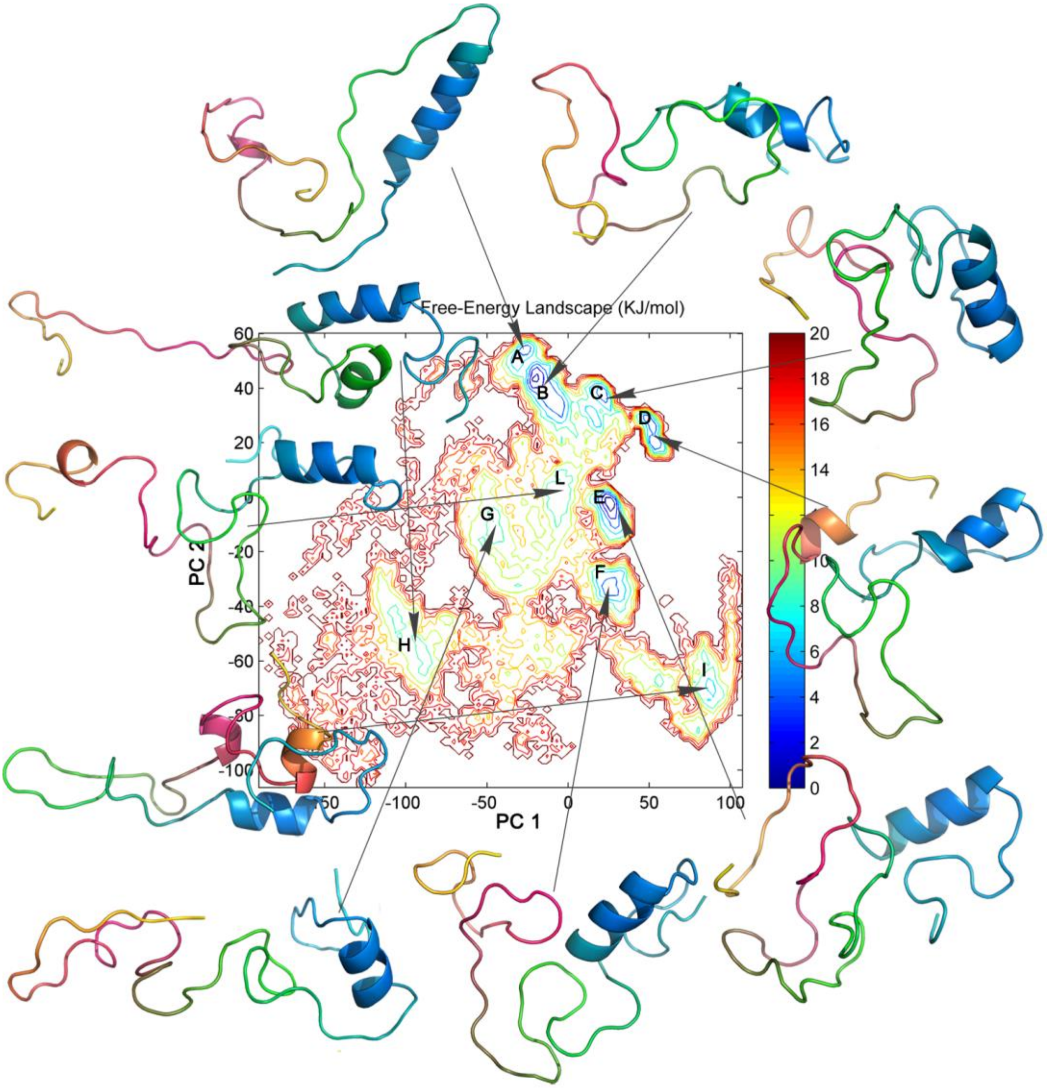


Figure 5. FEL representation calculated using the first two principal components as reaction coordinates. The major basins are labeled by capital letters (A to L). The free energy is given in kJ/mol and indicated by the color code shown on the figure. The average structure identified for each conformational basin by structural clustering are represented as cartoons. The average structures are highlighted by a color gradient, from N-terminus (cyan) to C-terminus (yellow).

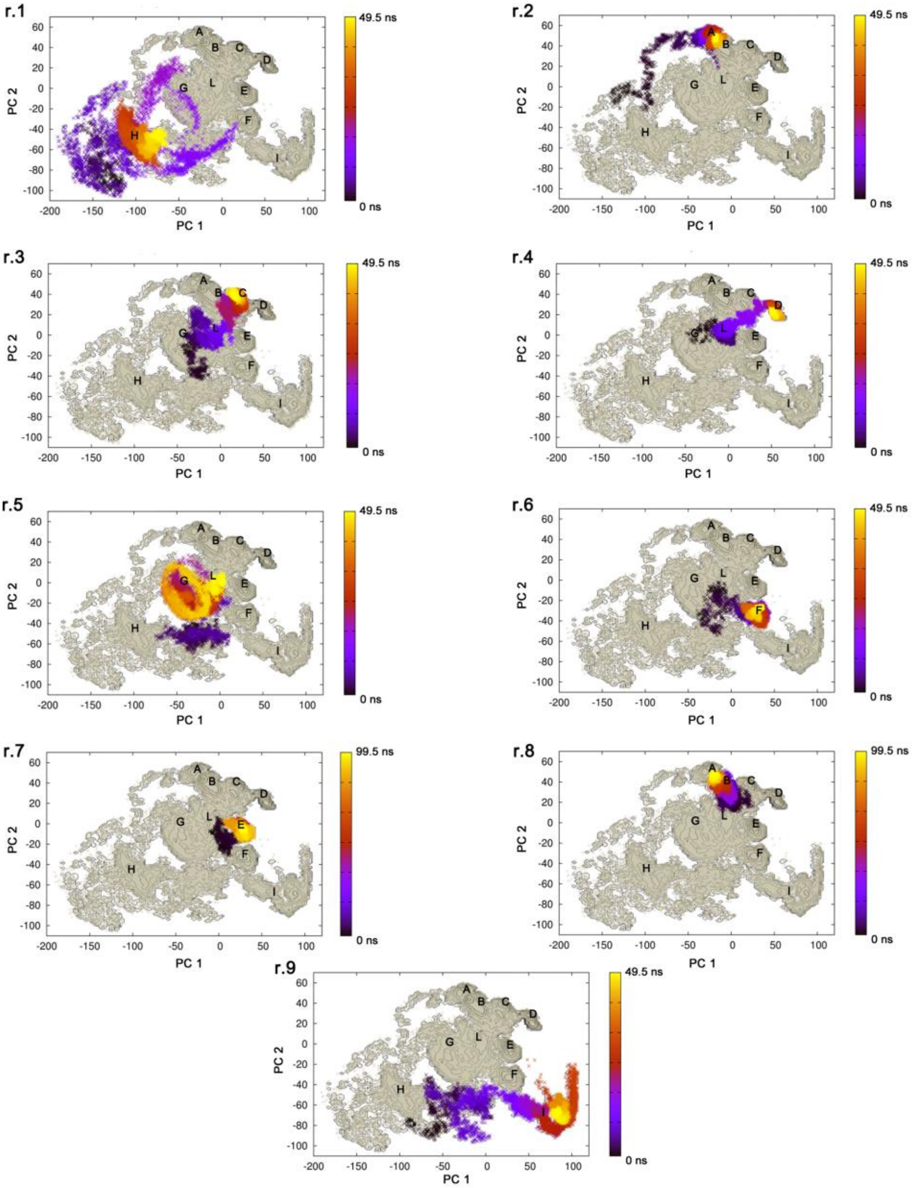


Figure 6. Projection of the single replicate on the FEL. The trajectories explored by each replicate are point out on the FEL representation obtained using the first two principal components as reaction coordinates. The temporal evolution of each replicate is indicated by a color gradient, from black (0 ns) to yellow (the last frame of replicate).

For each FEL basin, we have calculated the secondary-structure content, the SAS values, as well as the networks of salt-bridge, aromatic, amino-aromatic and hydrophobic interactions and their persistence (Tables 2 and 3). The average structures extracted from the ensemble trajectories have different structural properties, with less (A, G, H, L) and more compact (C, D, F, E) conformations, highlighting the intrinsic structural flexibility of Sic1 KID. Indeed, starting from completely extended models, the protein populates quite different conformational states, from extended random-coil-like conformations to highly compact structures. The here described principal motions of Sic1 KID identify distinct N- and C-terminal sub-domains. The first PC mainly describes the dynamics of the C-terminal region (yellow and hot-pink in the snapshots shown in Figure 5), whereas the second principal component mostly describes the motions of the N-terminal region (cyan in the snapshots shown in Figure 5). These two motions dictate the major features of the dynamic behavior of Sic1 KID, determining the pairing of the C-terminal and N-terminal regions and the transitions between an *open* and a *closed* state of the domain. The closed state is characterized by a long unstructured loop in the central region of the domain, approximately composed by the residues 243-270 (green in the snapshots shown in Figure 5).

Table 2. Secondary-structure content and SAS values. The secondary-structure content for each basin has been calculated considering all the different types of helical structures. Values are shown as number of residues or as percentage of the total number of residues in the Sic1 KID fragment. The minimum, maximum and average SAS values are displayed.

Basin name	*Helical content avg. (α-3_{10}-π)*	*Helical content max. (α -3_{10}- π)*	*Total helical content avg. %*	*SAS avg. (nm^2)*	*SAS min. (nm^2)*	*SAS max. (nm^2)*
A	16.42-0.01-1.23	23-6-7	27.15	60.10	53.45	67.63
B	7.35-1.89-0.94	22-7-10	15.71	56.32	49.70	70.87
C	10.65-0.09-0.08	16-6-6	17.14	53.57	47.84	68.94
D	13.2-0.01-0.12	16-5-4	20.00	54.71	50.05	60.05
E	10.54-0.16-0.11	16-5-6	15.71	55.94	50.05	68.32
F	7.33-1.97-0.06	21-11-6	14.28	55.31	48.70	68.81
G	10.89-0.02-0.12	22-8-6	17.14	58.72	53.10	75.83
H	17.75-0.01-0.03	21-5-6	25.71	65.24	60.25	71.26
I	14.51-0-0.23	24-0-6	22.85	57.58	53.08	64.29
L	12.11-0.07-0.26	15-9-6	18.57	60.78	53.87	71.74

Table 3. Hub residues identified for each conformational basin. The residues involved in three or more connections in the salt-bridge networks have been considered as hubs.

Basin	***Hub residues***
A	K260, E266
B	D243, E245, D246, K254, E256, E259, R261, R262, E266, E267, K268, R270
C	E 223, E240, D265, R270
D	R233, E245, R261, D265, E267, R270
E	R270, D281
F	K268
G	K234, K254, R261, R270, E283
H	R262, E266
I	R262, D265, K268, R269, D281
L	E240, R261, D265, R270

Secondary-structure content

The MD data were analysed to identify putative IFSUs, i.e. regions that are characterized by at least transient secondary structure during the simulation time. The secondary-structure content was calculated for each FEL basin and compared with experimental data obtained by FT-IR spectroscopy (Brocca et al., 2011b). The FT-IR spectra of Sic1 KID point out a high contribution of random-coil conformation (~ 40%) in addition to a ~ 30% of dynamic helical structures (α, 3_{10} and π helix). The average secondary-structure content calculated for each MD basin is reported in Table 2. The data extracted from MD simulations are in overall good agreement with FT-IR data, although slightly under-estimated. Such a discrepancy is likely due to inherent limits of the *GROMOS96* force field when sampling α-helical conformations (Matthes and de Groot, 2009). The structures belonging to basins A and H (Table 2) provide the best agreement with the available experimental data (Brocca et al., 2011b). The average structure derived from basin A displays a finger-like structure, composed by a long α–helix from residue L224 to residue R239. Instead, the average structure from basin H displays two α-helices, from Q227 to L238 and from I244 to T249.

Intramolecular interactions

To characterize the tertiary-structure properties of the distinct conformations populated by Sic1 KID, the SAS values were calculated for the different FEL basins (Table 2). The average SAS values range from 53 to 65 nm^2 (Table 2). These results can be compared with estimates obtained by ESI-MS performed in collaboration with Dr. L. Testa in the group of Prof. R. Grandori at the University of Milano-Bicocca since the extent of ionization correlates with the SAS of the protein at the moment of its transfer to the gas phase (Testa et al., 2011a).

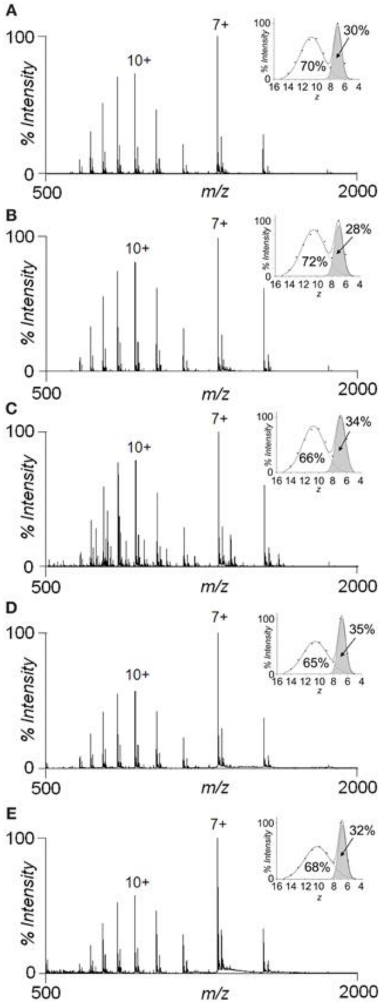


Figure 7. Species distributions by ESI-MS and effects of organic solvents. Nano-ESI-MS spectra of 10 µM protein in (A) 50 mM ammonium acetate, pH 6.5; (B) 50 mM ammonium acetate, pH 6.5, 30% acetonitrile; (C) 50 mM ammonium acetate, pH 6.5, 50% acetonitrile; (D) 50 mM ammonium acetate, pH 6.5, 30% methanol; (E) 50 mM ammonium acetate, pH 6.5, 50% methanol. The main charge state of each component is labeled by the corresponding charge state (7+ for the compact form and 10+ for the extended form). The insets show the gaussian fitting of the CSDs upon transformation to an x=z abscissa axis.

The Sic1 KID fragment gives rise to bimodal charge-state distributions (CSDs) by non-denaturing ESI-MS (Brocca et al., 2011b) (Figure 7A), indicating coexistence of compact and extended conformations (Kaltashov and Abzalimov, 2008). The predominant component represents a highly extended state, while the minor component (apparent relative amount ~30%) represents a compact state that disappears under denaturing conditions, as previously reported (Brocca et al., 2011b). The average SAS values derived from the ESI-MS data (Brocca et al., 2011b) are 59.78 nm^2 for the compact state and 88.76 nm^2 for the extended state.

Thus, the computational results compare quite well with the experimental data regarding the more compact form. This finding is in line with expectations from classical MD simulations, which are likely to capture conformational properties of collapsed forms, leaving, however, extended states quite unexplored. The *in-silico* SAS values identify two groups of different compactness within the collapsed state. One group comprises conformations that mostly derive from the energy basins B, C, D, E, F, with more compact and globular structures, characterized by the lowest SAS values (53-55 nm^2). The second group includes mainly structures sampled in the energy basins A, G, H, I, L. They display slightly less compact conformations, characterized by slightly higher average SAS values (58-65 nm^2). These differences might reflect structural heterogeneity within the collapsed state that could not be detected by CSD or ion-mobility analysis (Brocca et al., 2011b). Nevertheless, it cannot be ruled out that the group with lower SAS values could represent an artifact due to a bias for overcompaction of unfolded proteins in the current force fields (Click et al., 2010; Knott and Best, 2012). To better describe the SAS profiles of the Sic1 KID states in our MD ensemble, the FEL was also calculated using the SAS itself and the first and second PCs as reaction coordinates (Figures 8A,B).

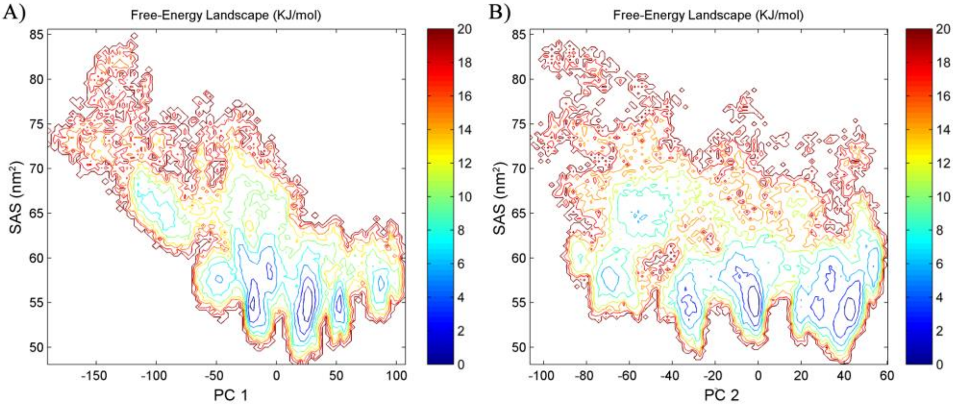


Figure 8. FEL representations calculated using the first (A) or the second (B) principal component and SAS values as reaction coordinates. The free energy is given in kJ/mol and indicated by the color code shown on the figure.

These FEL representations also point out the presence of several substates of different compactness. The most populated one is characterized by a SAS value around 55 nm^2 and corresponds to the compact conformations discussed above (from basins B, C, D, E, F). To investigate the driving force that likely promotes structural compaction in Sic1 KID, the different types of non-covalent interactions were examined for each FEL basin. In particular, salt-bridge, aromatic, hydrophobic and amino-aromatic interactions were analysed for persistence over the simulation time. This analysis identifies salt bridges as the major factor stabilizing compact structures, with a large number of basic and acidic residues involved in multiple interactions (Figure 9). As typical for IDPs, Sic1 KID has a low mean hydrophobicity. It also displays quite low net charge per residue (Brocca et al., 2011b). In spite of the presence of several charged residues, often consecutive in the amino acid sequence, these features suggest that Sic1 KID can be an IDP with propensity for globular states (Mao et al., 2010). Networks of salt bridges were reported to play a major role in IDP compaction, promoting formation of helical structures, and, in general, order/disorder transitions (Mao et al., 2010). In Sic1 KID, the salt-bridge networks have a transient nature and each conformational state is mainly stabilized by different pairs of interacting residues. Nevertheless, the analysis of salt-bridge networks in the MD ensemble identifies a subset of residues acting as hubs in the networks (Figure 10). They are likely to represent important residues in the development and maintenance of tertiary structure (Vishveshwara et al., 2009; Angelova et al., 2011). Despite the similar number of pairwise electrostatic interactions, the most compact conformations (basins B, C, D and E) have salt-bridge networks characterized by more highly interconnected residues than the less compact states (basins A, G, H and L) (Figures 9 and 10). The number of hub residues is generally greater for the more globular conformations (see for example structures from basins D and B) (Figure 10 and Table 3). Interactions involving hub residues in the less compact conformations have lower persistence than those in compact structures. These observations strengthen the importance of electrostatic interactions in the intrinsic structure of Sic1 KID. Some of the hub residues are shared by several compact structures, in particular R270, K268, E267, E245, R261 and D265. The analysis of the major paths connecting charged residues in the graph points out that hubs are also highly interconnected to each other in the compact states, showing multiple paths connecting them with high persistence values (Figure 10). Sub-networks of salt bridges have also been identified (Figure 11). Less compact states show a greater number of small size and poorly connected sub-networks. In fact, these are generally composed by isolated salt bridges or three/four-node networks. On the contrary, the globular states generally feature three major sub-networks, composed by a higher number of well-interconnected residues. These sub-networks include mainly residues from 240 to 280.

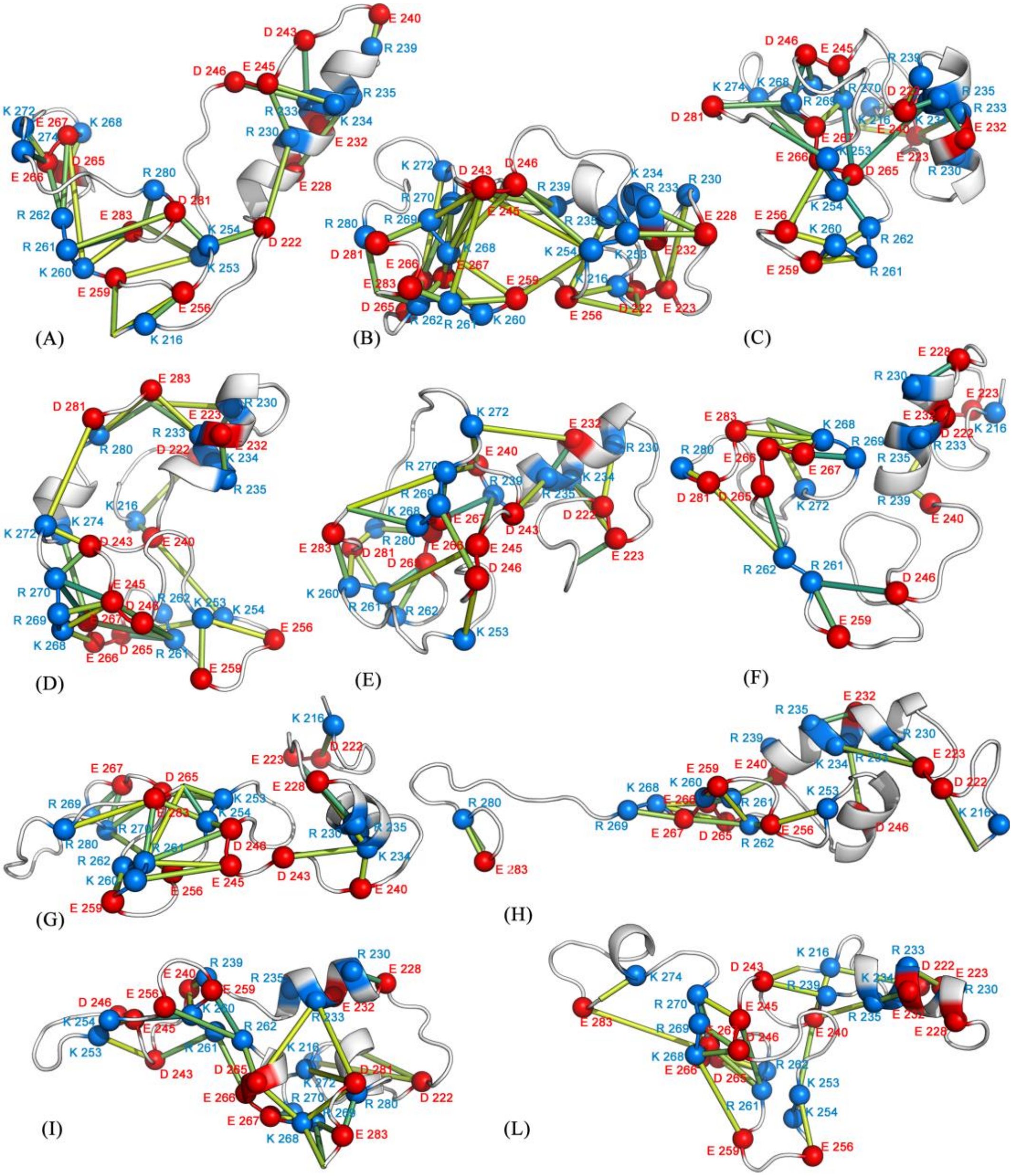


Figure 9. Networks of salt bridges. The average structures of each FEL basin are shown as cartoons, and the Cα atoms of the acidic and basic residues involved in salt bridges are shown as red and blue spheres, respectively. The Cα atoms of the interacting residues are connected by sticks of different shades of color depending on the interaction persistence (from yellow to green for increasing persistence values).

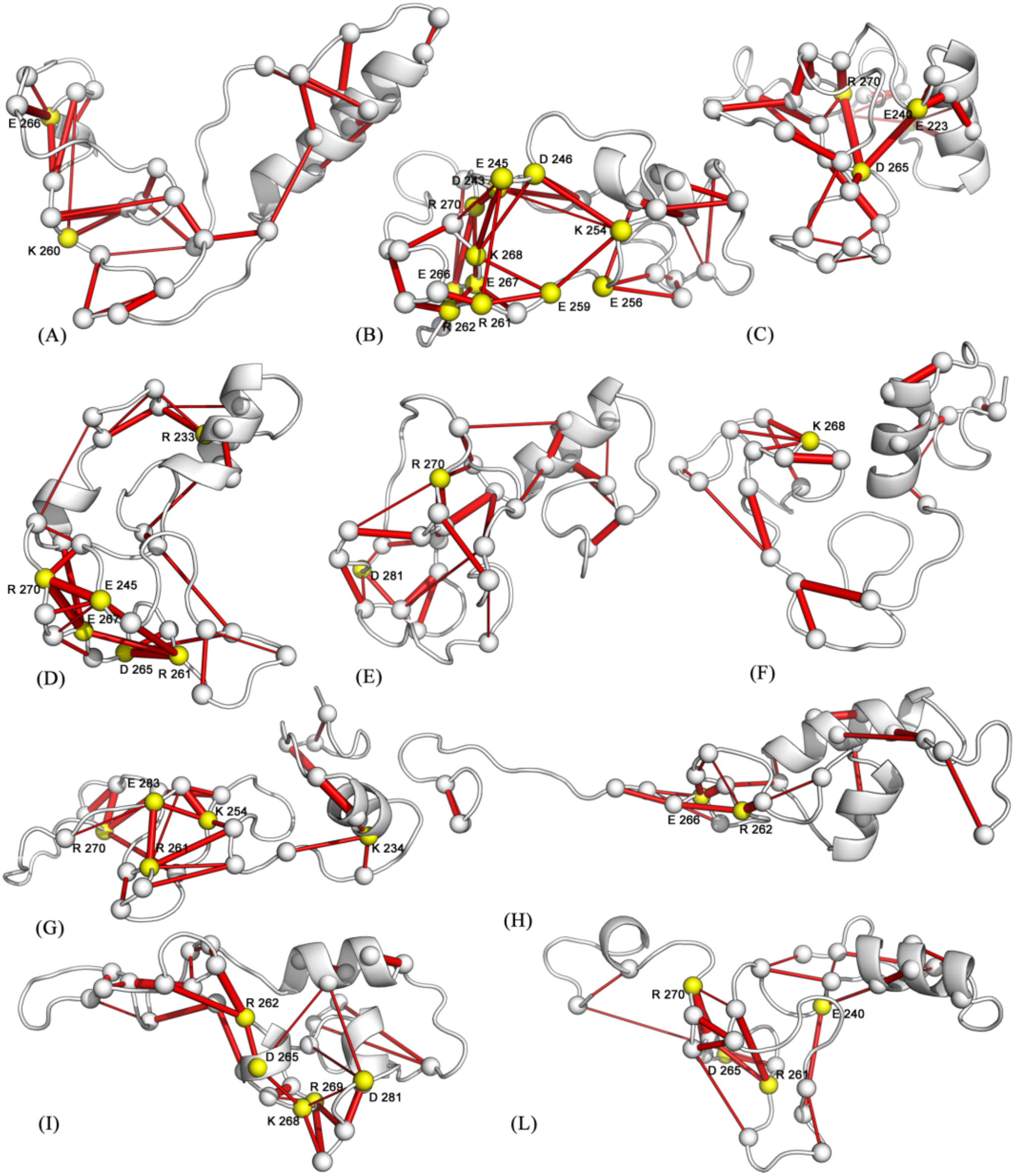


Figure 10. Hub residues in salt-bridge networks. The average structures of each conformational basin are shown as cartoons, and the Cα atoms of the residues involved in salt bridges are shown as spheres. The Cα atoms of the interacting residues are connected by sticks, whose thickness is proportional to the persistence of the interaction. Hub residues, defined as those involved in at least three different salt-bridge interactions, are highlighted in yellow.

Some amino-aromatic interactions could also be detected, even if generally very transient. Among the residues involved in amino-aromatic interactions with highest persistence, we can find Y250, F271, R239 and R235 (Figure 12). Aromatic-aromatic interactions are absent, with exception of basins B and I, where interactions involving F277 and other phenylalanines can be detected (*data not shown*). Hydrophobic-interaction networks are also present, principally in conformations derived from basins C and I, but they are characterized by very low persistence (*data not shown*). Moreover, an additional search for hydrophobic interactions, lowering the distance cutoffs from 0.5 to 0.45 nm causes a complete loss of these interactions. A marginal role of hydrophobic interactions in the stabilization of Sic1 KID compact structures is also indicated by ESI-MS experiments (Figure 7). The spectrum of the protein in 50 mM ammonium acetate, pH 6.5 is characterized by the aforementioned bimodal CSD, indicating coexistence of compact and extended states. The addition of acetonitrile or methanol up to 50% (the highest tested concentration) does not significantly affect the CSD, leaving clearly preserved the compact state. On the contrary, the compact form can be readily *denatured* by acids, as previously shown (Brocca et al., 2011b).

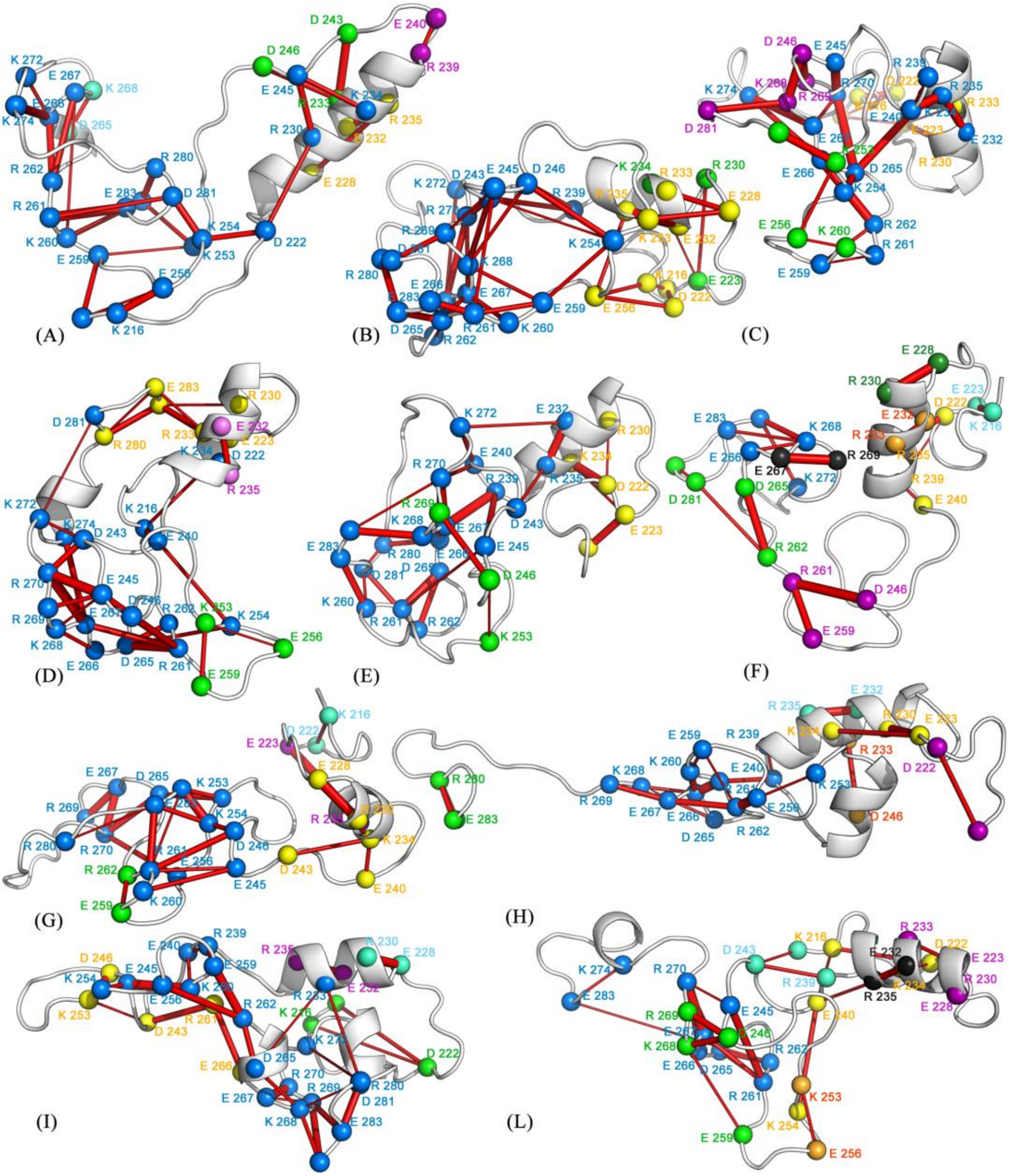


Figure 11. Sub-networks of salt-bridges. The average structures of each conformational basin are shown as cartoons, and the Cα atoms of the residues involved in salt bridges are shown as spheres. The Cα atoms of the interacting residues are connected by sticks, whose thickness is proportional to the persistence of the interaction. The single sub-networks are represented by a color code according to their size blue-yellow-green-purple-cyan-orange-black-dark green, going from the largest (21 residues) to smallest size (2 residues).

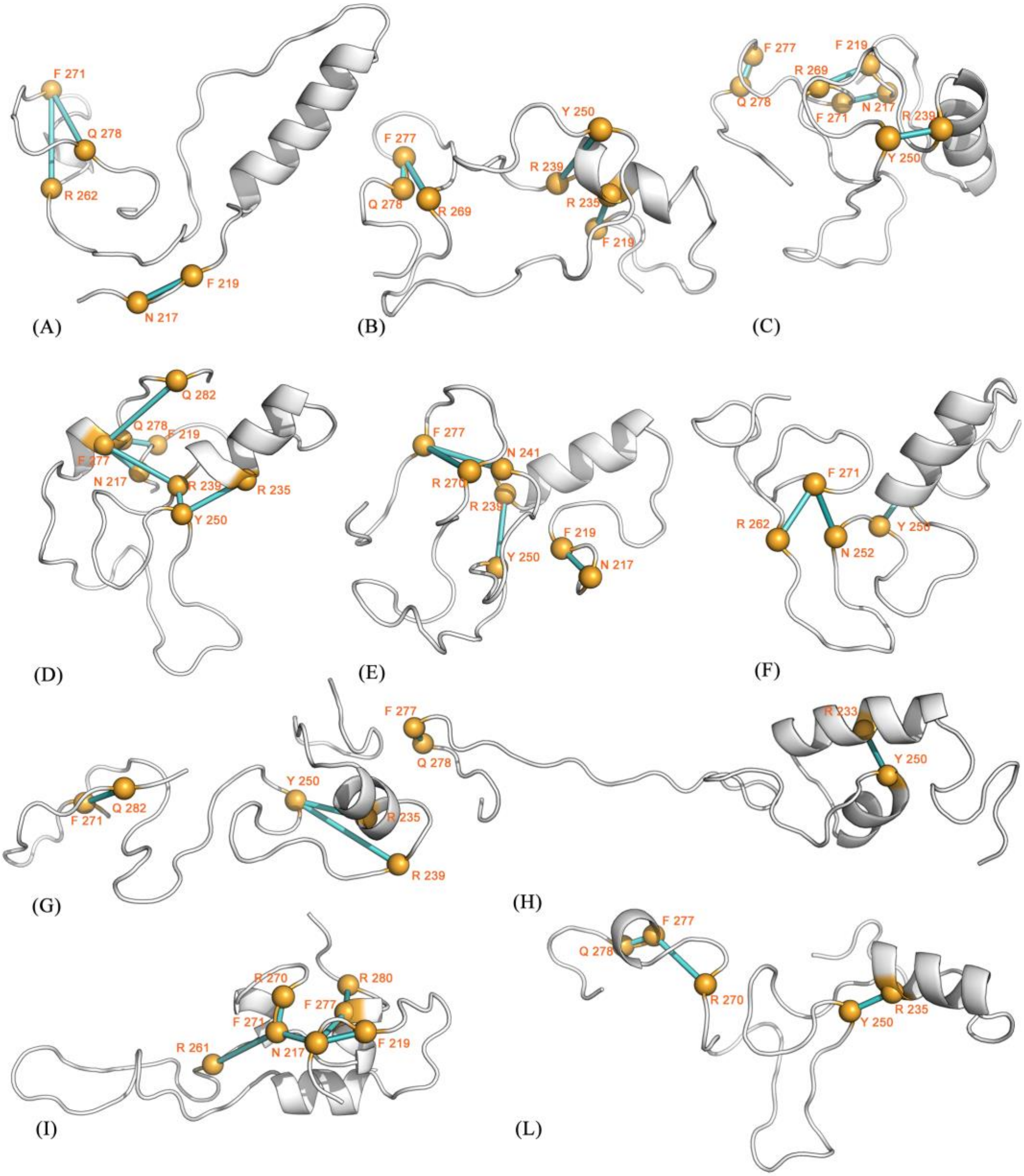


Figure 12. Amino-aromatic interactions. The average structures of each conformational basin are shown as cartoons, and the Cα atoms of the residues involved in amino-aromatic interactions are shown as yellow spheres. The Cα atoms of the interacting residues are connected by sticks of different shades of color, depending on the interaction persistence (from cyan to blue for increasing persistence).

2.2.4 Discussion

The secondary structure content and SAS values are in overall good agreement with experimental data obtained by FT-IR spectroscopy and ESI-MS pointing out coexistence of compact and extended conformations. The results point out that Sic1 KID has a highly dynamic behavior and strong propensity to structural disorder but can explore compact conformations, with considerable tertiary structure. The principal motions of Sic1 KID are related to the pairing of the C-terminal and N-terminal regions that are likely to represent the transitions between open and closed states. The different types of non-covalent interactions were examined, showing that electrostatic interactions are the major factor stabilizing compact structures as also indicated by ESI-MS experiments with a large number of basic and acidic residues involved in large networks. Residues acting as hubs in the intramolecular interaction networks have been identified, suggesting that they can have a key role in Sic1 architecture and function, likely to be involved in the stabilization of the globular states. Moreover, as for the unbound state of other IDPs, helical intrinsically folded structural units (IFSUs) have been identified: a persistent α-helix between E223 and L238, and a shorter transiently populated α-helix between residues I244 and I248. These transiently elements are thought to provide seeds of binding interfaces (Uversky, V. N. et al. *J Mol Recognit* 2005, 18, 343-384.) and be relevant to Sic1 function in vivo, might therefore promote interaction to the Cdk/cyclin complex in a mechanism similar to the one proposed for the mammalian p27 KID. The heterogeneous and dynamic nature of IDPs makes structural characterization of their unbound state highly challenging. Although MD force fields were developed to simulate protein folding, they proved useful to characterize the conformational ensembles of IDPs and unfolded proteins (Espinoza-Fonseca, 2009a; Cino et al., 2011; Arrigoni et al., 2012; Ganguly et al., 2012; Lindorff-Larsen et al., 2012; Knott and Best, 2012), especially when comparing computational results with biophysical data. These studies contributed to enforce the applicability of classical MD to complex molecular ensembles. Nevertheless, the study of dynamic and heterogeneous systems such as IDPs has to face limits in force field accuracy and sampling efficacy (Esteban-Martin et al., 2012). Thus, while helping description of globular IDP states, classical MD simulations are not adequate to describe the actual equilibrium between extended and compact conformations. This complementary information can be provided by experimental assessment of species distributions, for instance by MS (Kaltashov and Abzalimov, 2008) or NMR investigation (Esteban-Martin et al., 2012; Schneider et al., 2012). We here employ atomistic, explicit solvent, MD simulations integrated by experimental data (Brocca et al., 2011b) to provide a first atomic-level description of the conformational ensemble of compact states of the isolated Sic1 KID fragment. The results indicate that, in spite of its strong propensity for structural disorder, Sic1 KID can explore compact conformations, with considerable secondary and tertiary structure. The extents of secondary structure and solvent accessibility derived by the simulations are in good agreement with experimental results obtained by FT-IR spectroscopy and ESI-MS (Brocca et al., 2011b). The conformational ensemble of Sic1 KID reveals a highly dynamic behavior, populating several different conformations. Also local conformations, such as helical IFSUs are likely to be highly dynamic. Among tertiary contacts, electrostatic interactions are suggested to be the major determinants of structural compaction by both the ESI-MS and the MD results. This conclusion is consistent with the low mean hydropathy and the high mean net charge per residue of this protein (Brocca et al., 2009) and is in agreement with the current view on the importance of charged residues defining IDP structural and functional properties (Uversky et al., 2000). The present analysis also points out that the globular states of Sic1 KID are stabilized by highly interconnected networks of electrostatic interactions with a few hub residues in common to different conformations and involved in multiple paths. In particular R270, K268, E267, E245, R261 and D265 emerge as the most relevant ones. These residues represent good targets for mutagenesis experiments to further explore the role of such networks in Sic1 KID structure. Although our results do not hint to a major role of hydrophobic residues in intramolecular networks, they could still contribute to global compaction of the domain. The present results should also be interpreted in the light of the structural and functional relation to the mammalian p21 and p27 KID domains (Barberis et al., 2005). It has been shown that p27 can replace Sic1 in yeast cells (Barberis et al., 2005) and Sic1 KID can functionally interact with mammalian Cyclin A-Cdk2 and inhibit its kinase activity. The interaction between p27 and cyclin A/Cdk2 has been investigated suggesting a two-site, sequential binding process, in which p27 KID first interacts at one end with cyclin A (sub-domain D1), and then binds to Cdk2 by the other end (sub-domain D2), wrapping the central helical region (sub-domain LH) around the cyclin/kinase complex (Sivakolundu et al., 2005; Galea et al., 2008; Espinoza-Fonseca, 2009b; Otieno et al. 2011). The present results indicate the stretch between E223 and L238 as the most persistent α–helix of Sic1 KID, while a shorter and transiently populated α-helix approximately maps between residues I244 and I248. Although it is difficult to identify the exact boundaries of the helical regions by MD simulations of such a highly heterogeneous system, these regions of Sic1 correspond to the p27 LH sub-domain (residues 38-60), according to the structural alignment of the two KID domains (Barberis et al., 2005). This sub-domain has been identified as an IFSU also in p27 (Sivakolundu et al., 2005) and it is thought to play a role tethering the D1 to the D2 sub-domain and enhancing the overall ΔG (free energy) of binding (Otieno et al., 2011). The corresponding and structurally similar region of Sic1 KID might therefore promote binding to the Cdk/cyclin complex by a similar mechanism. More studies will be needed to test this hypothesis and further biochemical investigation will be needed to characterize the physiological intermediates of Sic1 binding. Furthermore, this conclusion is consistent with the low mean hydropathy and low mean net charge per residue of this protein (Brocca et al., 2009, 2011b) and is in agreement with the current view on the importance of charged residues defining IDP structural and functional properties (Uversky et al., 2000). This feature could be also relevant *in vivo*, in relation to the multiple phosphorylation events that regulate Sic1 interactions (Nash et al., 2001; Mittag et al., 2008; Koivomagi et al., 2011). By altering short- and long-range electrostatic interactions, phosphorylation could effectively modulate the conformational properties of this IDP, even far from the site of the modification. (Johnson and Lewis, 2001; Arrigoni et al., 2012). In particular, the present analysis points out that the globular states of Sic1 KID are stabilized by interconnected networks of electrostatic interactions with a few hub residues common to different conformations and involved in multiple paths. R270, K268, E267, E245, R261, and D265 emerge as the most relevant ones. These residues represent good targets for mutagenesis experiments to further explore the role of such networks in Sic1 KID structure. Although our results do not hint to a major role of hydrophobic residues in intramolecular networks, these could still contribute to global compaction of the domain. Further experiments will be necessary to investigate their structural role. In conclusion, the here provided experimental and computational evidence indicates that Sic1 KID, though highly disordered, can acquire transient secondary and tertiary structure populating compact conformations.

2.2.5 Acknowledgments

This work was supported by the Standard HPC Grant 2011 and 2012 from CASPUR to E.P., by a grant FAR (Fondo Ateneo per la Ricerca) of the Università Milano-Bicocca to R.G., L.DG., S.B. and by a doctoral fellowship “Fondazione Fratelli Confalonieri” to L.T. The authors would like to thank Lilia Alberghina for careful reading of the manuscript and her fruitful comments, Matteo Tiberti and Alberto Arrigoni for their precious suggestions, as well as Luis Andrè Baptista for technical help.

2.2.6 References

Amadei, A., Ceruso, M. A., and Nola, A. D. (1999). On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. *Proteins* 36, 419-424.

Amadei, A., Linssen, A. B., and Berendsen, H. J. (1993). Essential dynamics of proteins. *Proteins* 17, 412-425.

Angelova, K., Felline, A., Lee, M., Patel, M., Puett, D., and Fanelli, F. (2011). Conserved amino acids participate in the structure networks deputed to intramolecular communication in the lutropin receptor. *Cell. Mol. Life Sci.* 68, 1227-1239.

Arrigoni, A., Grillo, B., Vitriolo, A., De Gioia, L., and Papaleo, E. (2012). C-terminal acidic domain of ubiquitin-conjugating enzymes: A multi-functional conserved intrinsically disordered domain in family 3 of E2 enzymes. *J. Struct. Biol.* 178, 245-259.

Barberis, M. (2012). Sic1 as a timer of Clb cyclin waves in the yeast cell cycle - design principle of not just an inhibitor. *FEBS J., in press.*

Barberis, M., De Gioia, L., Ruzzene, M., Sarno, S., Coccetti, P., Fantucci, P., Vanoni, M., and Alberghina, L. (2005). The yeast cyclin-dependent kinase inhibitor Sic1 and mammalian p27Kip1 are functional homologues with a structurally conserved inhibitory domain. *Biochem. J.* 387, 639-647.

Barberis, M., Klipp, E., Vanoni, M., and Alberghina, L. (2007). Cell size at S phase initiation: an emergent property of the G1/S network. *PLoS Comput. Biol*. 3, e64.

Barberis, M., Linke, C., Adrover, M. A., Gonzalez-Novo, A., Lehrach, H., Krobitsch, S., Posas, F., and Klipp, E. (2012). Sic1 plays a role in timing and oscillatory behaviour of B-type cyclins. *Biotechnol. Adv.* 30, 108-130.

Belle, V., Rouger, S., Costanzo, S., Liquiere, E., Strancar, J., Guigliarelli, B., Fournel, A., and Longhi, S. (2008). Mapping alpha-helical induced folding within the intrinsically disordered C-terminal domain of the measles virus nucleoprotein by site-directed spin-labeling EPR spectroscopy. *Proteins* 73, 973-988.

Bernado, P., and Svergun, D. I. (2012). Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Mol. BioSys.* 8, 151-167.

Besson, A., Dowdy, S. F., and Roberts, J. M. (2008). CDK inhibitors: Cell cycle regulators and beyond. *Developmental Cell* 14, 159-169.

Brinda, K. V., and Vishveshwara, S. (2005). A network representation of protein structures: Implications for protein stability. *Biophys. J.* 89, 4159-4170.

Brocca, S., Samalikova, M., Uversky, V. N., Lotti, M., Vanoni, M., Alberghina, L., and Grandori, R. (2009). Order propensity of an intrinsically disordered protein, the cyclin-dependent-kinase inhibitor Sic1. *Proteins* 76, 731-746.

Brocca, S., Testa, L., Samalikova, M., Grandori, R., and Lotti, M. (2011a). Defining structural domains of an intrinsically disordered protein: Sic1, the cyclin-dependent kinase inhibitor of Saccharomyces cerevisiae. *Mol. Biotechnol.* 47, 34-42.

Brocca, S., Testa, L., Sobott, F., Samalikova, M., Natalello, A., Papaleo, E., Lotti, M., De Gioia, L., Doglia, S. M., Alberghina, L., and Grandori, R. (2011b). Compaction properties of an intrinsically disordered protein: Sic1 and its kinase-inhibitor domain. *Biophys. J.* 100, 2243-2252.

Brunger, A. T. (2007). Version 1.2 of the Crystallography and NMR system. *Nature Protocols* 2, 2728-2733.

Cino, E. A., Wong-ekkabut, J., Karttunen, M., and Choy, W. Y. (2011). Microsecond Molecular Dynamics Simulations of Intrinsically Disordered Proteins Involved in the Oxidative Stress Response. *Plos One* 6, e27371.

Click, T. H., Ganguly, D., and Chen, J. (2010). Intrinsically disordered proteins in a physics-based world. *Int. J. Mol. Sci.* 11**,** 5292-5309.

Coccetti, P., Rossi, R. L, Sternieri, F., Porro, D., Russo, G. L., Di Fonzo, A., Magni, F., Vanoni, M., and Alberghina, L. (2004). Mutations of the CK2 phosphorylation site of Sic1 affect cell size and S-Cdk kinase activity in Saccharomyces cerevisiae. *Mol. Microbiol.* 51, 447-460.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Clifford, S. (2009). Introduction to Algorithms. 3rd ed. Vol.1, 1292.

Darden, T., York, D., and Pedersen, L. (1993). Particle Mesh Ewald- an N.LOG(N) method for Ewald sums in large systems. *J. Chem. Phys*. 98, 10089-10092.

Deshaies, R. J., and Ferrell, J. E. Jr. (2001). Multisite phosphorylation and the countdown to S phase. *Cell* 107, 819-822.

Dobo, A., and Kaltashov, I. A. (2001). Detection of multiple protein conformational ensembles in solution via deconvolution of charge-state distributions in ESI MS. *Anal. Chem.* 73**,** 4763-4773.

Dunker, A. K., Silman, I., Uversky, V. N., and Sussman, J. L. (2008). Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.* 18, 756-764.

Dyson, H. J., and Wright, P. E. (2005). Intrinsically unstructured proteins and their functions.*Nature Rev. Mol. Cell Biol.* 6, 197-208.

Escote, X., Zapater, M., Clotet, J., and Posas, F. (2004). Hog1 mediates cell-cycle arrest in G1 phase by the dual targeting of Sic1. *Nat. Cell. Biol.* 6, 997-1002.

Espinoza-Fonseca, L. M. (2009a). Leucine-rich hydrophobic clusters promote folding of the N-terminus of the intrinsically disordered transactivation domain of p53. *FEBS Lett.* 583, 556-560.

Espinoza-Fonseca, L. M. (2009b). Reconciling binding mechanisms of intrinsically disordered proteins. *Biochem. Biophys. Res. Commun.* 382, 479-482.

Espinoza-Fonseca, L. M. (2012). Dynamic optimization of signal transduction via intrinsic disorder. *Mol. Biosyst*. 8, 194-197.

Espinoza-Fonseca, L. M., Kast, D., and Thomas, D. D. (2007). Molecular dynamics simulations reveal a disorder-to-order transition on phosphorylation of smooth muscle myosin. *Biophys. J.* 93, 2083-2090.

Espinoza-Fonseca, L. M., Kast, D., and Thomas, D. D. (2008). Thermodynamic and structural basis of phosphorylation-induced disorder-to-order transition in the regulatory light chain of smooth muscle myosin. *J. Am. Chem. Soc.* 130, 12208-12209.

Espinoza-Fonseca, L. M., Ilizaliturri-Flores, I., and Correa-Basurto, J. (2012). Backbone conformational preferences of an intrinsically disordered protein in solution. *Mol. Biosyst.* 8**,** 1798-1805.

Esteban-Martin, S., Fenwick, R. B., and Salvatella, X. (2012). Synergistic use of NMR and MD simulations to study the structural heterogeneity of proteins. *Wiley Interdiscipl. Rev. Comput. Mol. Sci.* 2, 466-478.

Eswar, N., Eramian, D., Webb, B., Shen, M. Y., and Sali, A. (2008). Protein structure modeling with MODELLER. *Methods Mol. Biol.* 426, 145-159.

Fisher, C. K., and Stultz, C. M. (2011). Constructing ensembles for intrinsically disordered proteins. *Curr. Op. Struct. Biol.* 21, 426-431.

Fuhrmans, M., Sanders, B. P, Marrink, S. J., and de Vries, A. H. (2010). Effects of bundling on the properties of the SPC water model. *Theor. Chem. Acc.* 125, 335-344.

Galea, C. A., Wang, Y., Sivakolundu, S. G, and Kriwacki, R. W. (2008). Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. *Biochemistry* 47, 7598-7609.

Ganguly, D., Zhang, W., and Chen, J. (2012). Synergistic folding of two intrinsically disordered proteins: searching for conformational selection. *Mol. BioSys.* 8, 198-209.

Gardebien, F., Thangudu, R. R., Gontero, B., and Offmann, B. (2006). Construction of a 3D model of CP12, a protein linker. *J. Mol. Graph. Model.* 25, 186-195.

Hazy, E., and Tompa, P. (2009). Limitations of Induced Folding in Molecular Recognition by Intrinsically Disordered Proteins. *Chemphyschem* 10, 1415-1419.

Hess, B. (2002). Convergence of sampling in protein simulations. *Phys. Rev. E Stat. Nonlin. Soft. Matter Phys.* 65, 031910.

Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. (1997). LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* 18, 1463-1472.

Hodge, A., and Mendenhall, M. (1999). The cyclin-dependent kinase inhibitory domain of the yeast Sic1 protein is contained within the C-terminal 70 amino acids. *Mol. Gen. Genet.* 262, 55-64.

Johnson, L. N., and Lewis, R. J. (2001). Structural basis for control by phosphorylation. *Chem. Rev.* 101**,** 2209-2242.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577-2637.

Kaltashov, I. A., and Abzalimov, R. R. (2008). Do ionic charges in ESI MS provide useful information on macromolecular structure? *J. Am. Soc. Mass Spectrom.* 19, 1239-1246.

Keller, B., Daura, X., and van Gunsteren, W. F. (2010). Comparing geometric and kinetic cluster algorithms for molecular simulation data. *J. Chem. Phys.* 132, 074110.

Kjaergaard, M., Teilum, K., and Poulsen, F. M. (2010). Conformational selection in the molten globule state of the nuclear coactivator binding domain of CBP. *Proc. Natl. Acad. Sci. U.S.A.* 107, 12535-12540.

Knott, M., and Best, R. B. (2012). A preformed binding interface in the unbound ensemble of an intrinsically disordered protein: evidence from molecular simulations. *PLoS Comput. Biol.* 8**,** e1002605.

Koivomagi, M., Valk, E., Venta, R., Iofik, A., Lepiku, M., Balog, E. R., Rubin, S. M., Morgan, D. O., and Loog, M. (2011). Cascades of multisite phosphorylation control Sic1 destruction at the onset of S phase. *Nature* 480**,** 128-131.

Lindorff-Larsen, K., Piana, S., Dror, R. O., and Shaw, D. E. (2011). How fast-folding proteins fold. *Science* 334**,** 517-520.

Lindorff-Larsen, K., Trbovic, N., Maragakis, P., Piana, S., and Shaw, D. E. (2012). Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *J. Am. Chem. Soc.* 134**,** 3787-3791.

Lopez-Aviles, S., Kapuy, O., Novak, B., and Uhlmann, F. (2009). Irreversibility of mitotic exit is the consequence of systems-level feedback. *Nature* 459**,** 592-595.

Mao, A. H., Crick, S. L., Vitalis, A., Chicoine, C. L., and Pappu, R. V. (2010). Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U.S.A.* 107**,** 8183-8188.

Marsh, J. A., Neale, C., Jack, F. E., Choy, W. Y., Lee, A. Y., Crowhurst, K. A., and Forman-Kay, J. D. (2007). Improved structural characterizations of the drkN SH3 domain unfolded state suggest a compact ensemble with native-like and non-native structure. *J. Mol. Biol.* 367**,** 1494-1510.

Matthes, D., and De Groot, B. L. (2009). Secondary Structure Propensities in Peptide Folding Simulations: A Systematic Comparison of Molecular Mechanics Interaction Schemes. *Biophys. J.* 97**,** 599-608.

Mendenhall, M. D., and Hodge, A. E. (1998). Regulation of Cdc28 cyclin-dependent protein kinase activity during the cell cycle of the yeast Saccharomyces cerevisiae. *Microbiol. Mol. Biol. Rev.* 62**,** 1191-1243.

Meszaros, B., Simon, I., and Dosztanyi, Z. (2011). The expanding view of protein-protein interactions: complexes involving intrinsically disordered proteins. *Phys. Biol.* 8.

Mittag, T., Orlicky, S., Choy, W. Y., Tang, X., Lin, H., Sicheri, F., Kay, L. E., Tyers, M., and Forman-Kay, J. D. (2008). Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. *Proc. Natl. Acad. Sci. U.S.A.* 105**,** 17772-17777.

Morin, B., Bourhis, J. M., Belle, V., Woudstra, M., Carriere, F., Guigliarelli, B., Fournel, A., and Longhi, S. (2006). Assessing induced folding of an intrinsically disordered protein by site-directed spin-labeling electron paramagnetic resonance spectroscopy. *J. Phys. Chem. B* 110**,** 20596-20608.

Nash, P., Tang, X., Orlicky, S., Chen, Q., Gertler, F. B., Mendenhall, M. D., Sicheri, F., Pawson, T., and Tyers, M. (2001). Multisite phosphorylation of a CDK inhibitor sets a threshold for the onset of DNA replication. *Nature* 414**,** 514-521.

Norholm, A.-B., Hendus-Altenburger, R., Bjerre, G., Kjaergaard, M., Pedersen, S. F., and Kragelund, B. B. (2011). The Intracellular Distal Tail of the Na(+)/H(+) Exchanger NHE1 Is Intrinsically Disordered: Implications for NHE1 Trafficking. *Biochemistry* 50**,** 3469-3480.

Oldfield, C. J., Cheng, Y., Cortese, M. S., Brown, C. J., Uversky, V. N., and Dunker, A. K. (2005). Comparing and combining predictors of mostly disordered proteins. *Biochemistry* 44**,** 1989-2000.

Otieno, S., Grace, C. R., and Kriwacki, R. W. (2011). The role of the LH subdomain in the function of the Cip/Kip cyclin-dependent kinase regulators. *Biophys. J.* 100**,** 2486-2494.

Papaleo, E., Mereghetti, P., Fantucci, P., Grandori, R., and De Gioia, L. (2009). Free-energy landscape, principal component analysis, and structural clustering to identify representative conformations from molecular dynamics simulations: the myoglobin case. *J. Mol. Graph. Model.* 27**,** 889-899.

Pasi, M., Tiberti, M., Arrigoni, A., and Papaleo, E. (2012). xPyder: a PyMOL plugin to analyze coupled residues and their networks in protein structures. *J. Chem. Inf. Model, in press.*

Qin, Z., Kalinowski, A., Dahl, K. N., and Buehler, M. J. (2011). Structure and stability of the lamin A tail domain and HGPS mutant. *J. Struct. Biol.* 175**,** 425-433.

Rauscher, S., and Pomes, R. (2010). Molecular simulations of protein disorder. *Biochem. Cell Biol.* 88**,** 269-290.

Receveur-Bréchot, V., Bourhis, J. M., Uversky, V. N., Canard, B., and Longhi, S. (2006). Assessing protein disorder and induced folding. *Proteins* 62**,** 24-45.

Russo, A. A., Jeffrey, P. D., Patten, A. K., Massague, J., and Pavletich, N. P. (1996). Crystal structure of the p27Kip1 cyclin-dependent-kinase inhibitor bound to the cyclin A-Cdk2 complex. *Nature* 382**,** 325-331.

Salmon, L., Nodet, G., Ozenne, V., Yin, G., Jensen, M. R., Zweckstetter, M., and Blackledge, M. (2010). NMR characterization of long-range order in intrinsically disordered proteins. *J. Am. Chem. Soc.* 132**,** 8407-8418.

Schneider, R., Huang, J.-R., Yao, M., Communie, G., Ozenne, V., Mollica, L., Salmon, L., Jensen, M. R., and Blackledge, M. (2012). Towards a robust description of intrinsic protein disorder using nuclear magnetic resonance spectroscopy. *Mol. Biosyst.* 8**,** 58-68.

Schwob, E., Bohm, T., Mendenhall, M. D., and Nasmyth, K. (1994). The B-type cyclin kinase inhibitor p40SIC1 controls the G1 to S transition in S. cerevisiae. *Cell* 79**,** 233-244.

Sivakolundu, S. G., Bashford, D., and Kriwacki, R. W. (2005). Disordered p27(Kip1) exhibits intrinsic structure resembling the Cdk2/cyclin A-bound conformation. *J. Mol. Biol.* 353**,** 1118-1128

Szasz, C., Alexa, A., Toth, K., Rakacs, M., Langowski, J., and Tompa, P. (2011). Protein Disorder Prevails under Crowded Conditions. *Biochemistry* 50**,** 5834-5844.

Testa, L., Brocca, S., and Grandori, R. (2011a). Charge-Surface Correlation in Electrospray Ionization of Folded and Unfolded Proteins. *Anal. Chem.* 83**,** 6459-6463.

Testa, L., Brocca, S., Samalikova, M., Santambrogio, C., Alberghina, L., and Grandori, R. (2011b). Electrospray ionization-mass spectrometry conformational analysis of isolated domains of an intrinsically disordered protein. *Biotechnol. J.* 6**,** 96-100.

Tiberti, M., and Papaleo, E. (2011). Dynamic properties of extremophilic subtilisin-like serine-proteases. *J. Struct. Biol.* 174**,** 69-83.

Tompa, P. (2005). The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.* 579**,** 3346-3354.

Tompa, P., and Fuxreiter, M. (2008). Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem. Sci.* 33**,** 2-8.

Toyoshima, H., and Hunter, T. (1994). p27, a novel inhibitor of G1 cyclin-Cdk protein kinase activity, is related to p21. *Cell* 78**,** 67-74.

Turoverov, K. K., Kuznetsova, I. M., and Uversky, V. N. (2010). The protein kingdom extended: Ordered and intrinsically disordered proteins, their folding, supramolecular complex formation, and aggregation. *Prog. Biophys. Mol. Biol.* 102**,** 73-84.

Uversky, V. N. (2002). Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* 11**,** 739-756.

Uversky, V. N. (2011). Intrinsically disordered proteins may escape unwanted interactions via functional misfolding. *Biochim. Biophys. Acta* 1814**,** 693-712.

Uversky, V. N., and Dunker, A. K. (2010). Understanding protein non-folding. *Biochim. Biophys. Acta* 1804**,** 1231-1264.

Uversky, V. N., Gillespie, J. R., and Fink, A. L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* 41**,** 415-427.

Verma, R., Feldman, R. M., and Deshaies, R. J. (1997). SIC1 is ubiquitinated in vitro by a pathway that requires CDC4, CDC34, and cyclin/CDK activities. *Mol. Biol. Cell* 8**,** 1427-1437.

Vishveshwara, S., Ghosh, A., and Hansia, P. (2009). Intra and Inter-Molecular Communications Through Protein Structure Network. *Curr. Prot. Pept. Sci.* 10**,** 146-160.

Wostenberg, C., Kumar, S., Noid, W. G., and Showalter, S. A. (2011). Atomistic Simulations Reveal Structural Disorder in the RAP74-FCP1 Complex. *J. Phys. Chem. B* 115, 13731-13739.

Wright, P. E., and Dyson, H. J. (2009). Linking folding and binding. *Curr. Opin. Struct. Biol.* 19**,** 31-38.

Yaakov, G., Duch, A., Garcia-Rubio, M., Clotet, J., Jimenez, J., Aguilera, A., and Posas, F. (2009). The stress-activated protein kinase Hog1 mediates S phase delay in response to osmostress. *Mol. Biol. Cell* 20**,** 3572-3582.

Yoon, M.-K., Venkatachalam, V., Huang, A., Choi, B.-S., Stultz, C. M., and Chou, J. J. (2009). Residual structure within the disordered C-terminal segment of p21(Waf1/Cip1/Sdi1) and its implications for molecular recognition. *Protein Sci.* 18**,** 337-347.

Zhuravlev, P. I., Materese, C. K., and Papoian, G. A. (2009). Deconstructing the Native State: Energy Landscapes, Function, and Dynamics of Globular Proteins. *J. Phys. Chem. B* 113**,** 8800-8812.

2.3 The conformational ensemble of the disordered and aggregation-protective 182-291 region of ataxin-3.

Invernizzi, G., Lambrughi, M., Regonesi, M.E., Tortora, P., Papaleo, E. (2013).
Biochim. Biophys. Acta. :5236-47

2.3.1 Introduction

Many IDPs or proteins containing disordered domains are involved in a number of diseases, such as cancer, neurodegenerative and amyloidosis diseases (Uversky, 2009, Uversky et al., 2008). Representative of this class of multi-domain proteins are polyglutamine proteins, a unique class of polypeptides with different functions and sequences that share the distinctive feature of a tract of consecutive glutamines located in disordered regions of the protein (Zoghbi and Orr, 2000). These multi-domain proteins have been extensively investigated since they trigger amyloid-related neurodegeneration upon expansion of the polyQ tract (Scherzinger et al., 1997). Indeed the poly-Q diseases are neurodegenerative hereditable pathologies associated with aberrant aggregation and formation of amyloid aggregates in different region of the brain (Tompa, 2005). Notably, it has been demonstrated that sequences flanking the polyQ can deeply influence the aggregation pathway and consequently the toxicity of the protein (Nozaki et al., 2001, Duennwald et al., 2006, Ellisdon et al., 2006, Thakur et al., 2009). Furthermore, it has been shown that early steps of the aggregation process are polyQ independent for ataxin-3 (AT3) and Huntingtin (Masino et al., 2003, Mao et al., 2005). AT3 is one of the best characterized polyQ proteins, composed by a folded N-terminal domain and a disordered C-terminal part. The latter harbors the polyQ and two ubiquitin (Ub) interaction motifs (UIMs) necessary for the AT3 deubiquitinase activity (Masino et al., 2003, Burnett et al., 2003, Nicastro et al., 2010). The AT3 N-terminal folded domain, the so-called Josephin domain (JD) is the only AT3 domain for which structural information is available in atomic details (Masino et al., 2004, Chow et al., 2004). JD is amyloidogenic on its own and it is responsible for the initial steps of AT3 aggregation (Masino et al., 2003). Recently, we have demonstrated that the disordered part spanning from JD to polyQ (tract 182–291) increases the aggregation rate of JD and gives rise to aggregates with structural properties different from those of the isolated JD (Santambrogio et al., 2012). Furthermore, it has been demonstrated that an N-terminal truncated AT3 variant at residue 259, which is physiologically relevant as a cleavage product of caspases, is toxic and induces an ataxic phenotype in a murine model (Hubener et al., 2011).

In light of the above observations, the disordered tract between the JD and the polyQ seems to play a key role in the aggregation process and related toxicity of AT3. Interestingly, the disordered tract does not harbor any predicted aggregation prone region (Santambrogio et al., 2012) nor does its presence alter the structural stability of the JD (Masino et al., 2004, Chow et al., 2004), leaving the open question of how it could influence JD aggregation. To shed light on this fundamental issue, an atomistic description of the conformational ensemble of the whole N-terminal part of AT3 is required to address the influence of the disordered tract on JD structural properties. The structural characterization of the disordered part thus becomes a mandatory step towards this goal.

Herein, we provide the first model of the conformational ensemble of the 182–291 region of AT3 in solution by atomistic explicit solvent multi-replicate simulations integrated to experimental data achieved by Circular Dichroism (CD), Size-Exclusion Chromatography (SEC) and ElectroSpray Ionization Mass Spectrometry (ESI-MS). It turns out to be an intrinsically disordered domain, characterized by α-helix structures in correspondence with the UIMs, as well as in other surrounding regions. The domain is likely to populate both poorly compact conformations with few long range intramolecular interactions and globular states mainly maintained by electrostatic interactions, most of which involve UIM elements.

2.3.2 Materials and Methods

Analysis of the primary sequence

The net charge per residue was calculated according to Mao et al. (Mao et al., 2010). Sequence hydrophobicity was calculated by the Kyte and Doolittle approximation by a window size of 5 amino acids and normalized in a scale between 0 and 1. The mean net charge per residue and mean hydrophobicity values were used to discriminate among intrinsically disordered and natively folded proteins, as proposed by Uversky et al. (Uversky et al., 2000).

Molecular dynamics simulations

MD simulations were performed by Gromacs 4 (www.gromacs.org), implemented on a parallel architecture, with the GROMOS96 43a1. The starting models of the $AT3_{182–291}$ for MD simulations were generated by ab-initio modeling by I-Tasser (http://zhanglab.ccmb.med.umich.edu/I-TASSER/(Zhang, 2008)), also including the NMR structure of the AT3 UIMs in the Ub-bound conformation (PDB ID: 2KLZ) as one of the references. 35 models were generated, including distance restraints between residues L233-I253, D241-R250, I240-L249 (distances higher than 2 nm) to guide I-TASSER modeling and to avoid artificial intramolecular interactions in the models. We then selected as starting structure for the simulations one of the models that lacks sidechain–sidechain long-range intramolecular interactions. The model was soaked in a triclinic box of 19622 Simple Point Charge (SPC) water molecules with periodic boundary conditions. The box was built so that all the protein atoms were at a distance of at least 1.5 nm from the box edges. Starting from this system of 60123 atoms (protein plus water atoms) we carried out a first preparatory 50 ns MD simulation that was not included in the final ensemble. The final structure after 50 ns was used as a starting conformation for the subsequent MD simulations. In these MD simulations, the box was resized so that all the protein atoms were at a distance of at least 0.8 nm from the box edges to speed up the simulations (i.e. the total system, including both protein and water atoms was of 23586 atoms). In fact, after the first preparatory run the gyration radius of the protein decreases, as discussed in the Results and discussion section. Thus, we could afford to carry out the simulations using a lower number of water molecules. Productive MD simulations were performed in the isothermal–isobaric ensemble (300 K, 1 bar). The LINCS algorithm was employed to constrain heavy atom bond lengths, allowing for the use of a 2 fs time-step. Long-range electrostatic interactions were calculated using Particle-Mesh Ewald (PME) summation scheme. Van der Waals and Coulomb interactions were truncated at 1.2 nm, accordingly to cutoffs previously applied for simulations of IDPs and experimentally validated by comparison with electronic paramagnetic resonance and fluorescence data (Espinoza-Fonseca et al., 2008, Espinoza-Fonseca, 2009). Na^+ and Cl^- counterions were added to the system to neutralize the overall charge and to simulate a physiological ionic strength (150 mM), according to a protocol previously employed for IDPs (Espinoza-Fonseca, 2009). The non-pair list was updated every 10 steps and conformations were stored every 4 ps.

8 unconstrained MD simulations were carried out over 50 ns, extended to 100 ns in ad hoc selected cases to verify trajectories convergence, achieving overall more than 0.50 μs of MD ensemble overall. The time evolution of the main chain root mean square (rmsd) deviation with respect to the starting structure for the 8 MD runs was used to assess stability of each trajectory. The first 10 ns of each replicate, that features a major drift in main chain rmsd, were not included in the further analyses (*data not shown*). After evaluation of the achieved conformational sampling by principal components analysis (see below), the remaining 40 ns of each replicate was joined in a macro-trajectory for further analyses. In fact, the first four principal components of the macro-trajectory accounts for more than 70% of the *essential subspace*.

Analyses of the simulations

The solvent accessible surface (SAS) was calculated by g_sas Gromacs tool. The secondary structure (ss) content was calculated by DSSP program and g_helix Gromacs tool, along with a residue-dependent persistence degree of ss profile (pdssp).

Salt-bridges and their networks, along with networks of aromatic, amino-aromatic and hydrophobic interactions, were analyzed employing a persistence cut-off of 20% and a distance cutoff of 0.5 nm. Aromatic interactions were also verified at a higher cutoff of 0.6 nm. Hydrogen bonds were analyzed using a persistence cut-off of 20%, a distance cutoff between donor and acceptor group of 0.3 nm and a minimum donor–H-acceptor angle of 120°. Salt-bridge networks were analyzed by the Pymol plugin xPyder (Pasi et al., 2012). xPyder represents pairwise relationships related to protein structures as two-dimensional matrices. In particular, the module for network analysis implemented in xPyder was employed. A network is described as a set of points (nodes) and connections between them (edges). A path is defined as a sequence of nodes for which an edge always exists between two consecutive nodes of the path. A matrix describing the persistence of each class of interactions was used as an input file. The program represents each residue of the matrix as a node of a simple, weighted graph connected by edges, whose weights are defined by the persistence of the interaction in the MD ensemble (i.e. the number of trajectory frames in which the interaction was present, over the total number of frames).

Residues connected by more than 3 edges to their neighbors are referred as hubs of the interaction network. The connected components, also called sub-networks, of the graph were also calculated. These are isolated sub-graphs in which all the edges are linked by at least one path, but no path exists between the nodes of the connected component and the rest of the graph. This analysis allows us to identify different clusters of interaction networks. The searching procedure was carried out so that the same node is not visited more than once to avoid entrapment in cycles.

Principal component analysis (PCA) and free energy landscape (FEL)

PCA highlights high-amplitude, concerted motions in MD trajectories, through the diagonalization of the mass-weighted covariance matrix (C) of the atomic positional fluctuations (Amadei et al., 1993). In particular, PCA analysis of MD trajectories provides a set of eigenvectors, each defined by an eigenvalue, describing the direction and the amplitude of the motion respectively. The first eigenvector represents the largest part of the total fluctuation of the system, the second eigenvector the second largest and so forth. The PCA calculations were performed both on the Cα carbons and all atoms coordinates of the $AT3_{182–291}$ structural ensemble generated by the macro-trajectory.

Given a reaction coordinate q_α, the probability of finding the system in a particular state q_α is proportional to ($e^{-G(q\alpha)/kT}$), where $G(q_\alpha)$ is the Gibbs free energy of that state. The FEL can be computed from the equation$G(q_\alpha) = -kTln[P(q_\alpha)]$, where k is the Boltzmann constant, T is the temperature of the simulation and $P(q_\alpha)$ is an estimation of the probability density function obtained from a histogram of the MD data. Considering two different reaction coordinates, for example q and p, the two-dimensional FEL can be obtained from the joint probability distribution $P(q,p)$ of these variables. In particular, the reaction coordinates considered in this study were the first three cartesian principal components (PCs or eigenvectors), as well as radius of gyration (Rg) and solvent accessible surface (SAS) values.

The conformational sampling was assessed by the root mean square inner product (rmsip) as an index of similarity between the essential subspaces obtained from the different independent simulations defined by their basis vectors (Amadei et al., 1999):

$$\mathbf{RMSIP} = \frac{1}{D}\sum_{i=1}^{D}\sum_{j=1}^{D}(\boldsymbol{\eta}_i^A\,\boldsymbol{\eta}_j^B)$$

Where $\boldsymbol{\eta iA}$ and $\boldsymbol{\eta jB}$ are the *i*th and *j*th eigenvectors from a set of eigenvectors A and B to be compared, and D is the number of eigenvectors considered. The rmsip was computed on the first 10 eigenvectors (Amadei et al., 1999).

Cluster analysis

The structures belonging to each main basin on the FEL were isolated and the main chain root mean square deviation (rmsd) matrix of each basin was calculated. The Gromos algorithm (Keller et al., 2010) was employed for clustering, using a cutoff of 0.4 nm. For each cluster, the structure with the lowest rmsd compared to the other cluster members was selected as the average structure.

Order parameter O

The order parameter *O* was calculated according to Chap. 2.2. The results obtained for AT3$_{182-291}$ by using the different cutoffs are reported in Table 1, along with the results calculated for the folded Josephin domain as a comparison.

Protein expression and purification

$AT3_{182–291}$ was cloned in fusion with glutathione S-transferase (GST) in a pGEX-6P-1 (GE Healthcare LifeSciences, Little Chalfont, England) plasmid and expressed in *E. coli* BL21 Codon Plus strain (Stratagene, La Jolla, CA, USA) as previously described [20]. Soluble protein fractions were purified on a Glutathione Sepharose 4 Fast Flow column and subsequently *in-site* cleaved with PreScission Protease® (GE Healthcare LifeSciences, Little Chalfont, England). The eluted samples were further purified by size-exclusion chromatography on a Superose 12 10/300GL column (GE Healthcare LifeSciences, Little Chalfont, England) in PBS buffer with a flow rate of 0.5 ml/min.

Circular dichroism

CD analyses were performed on a J-815 spectropolarimeter (JASCO corporation, Tokyo, Japan) in the far-UV region (190–250 nm) with 0.2 nm data pitch, accumulating ten spectra for each sample analysed. To obtain secondary structure content, spectra deconvolution was performed with Selcon3, CDSSTR and Continll algorithms in the CDpro package, with protein set 7.

Size-exclusion chromatography (SEC)

Size-exclusion chromatography was performed on an AKTA purifier liquid-chromatography system, using a Superose 12 10/300GL column (GE Healthcare LifeSciences, Little Chalfont, England). Chromatography was carried out in PBS buffer, pH 7.2, 150 mM NaCl at a flow rate of 0.5 mL/min, and monitored by absorbance at 280 nm. A calibration curve was derived by plotting the logarithm of the gyration radii of a set of standard proteins against their relative Kd values. Standards employed at 1.00 mg/mL: carbonic anhydrase II (29.1 kDa), bovine β-lactoglobulin (36.5 kDa), bovine aprotinin (6.5 KDa), bovine ribonuclease A (13.7 kDa), bovine cytochrome C (13 kDa) and blue dextran (2000 kDa) (Sigma Aldrich, St. Louis, MO).

ESI-MS

ESI-MS experiments were performed on a hybrid quadrupole-Time-of-Flight (q-TOF) mass spectrometer (QSTAR ELITE, Applied Biosystems, Foster City, CA, USA) equipped with a nano-electrospray ionization sample source. Metal-coated borosilicate capillaries (Proxeon, Odense, Denmark) with medium-length emitter tip of 1-μm internal diameter were employed. Instrumental parameters were: curtain gas 20 psi; ion spray voltage 1.1–1.4 kV; declustering potential 80 V. Analyses under non-denaturing conditions were performed in 10 mM ammonium acetate pH 6.5. Measurements at high temperature were performed with the interface heater set at 175 °C. Otherwise, the interface heater was switched off and the samples were sprayed at room temperature. All the reported spectra are averaged over 1-minute acquisition time.

2.3.3 Results and discussion

182–291 has no aggregation propensity by itself and exerts a protective role on aggregation of AT3

In a previous work, our groups showed that AT3$_{182–291}$ greatly enhances the aggregation kinetics of the JD (Santambrogio et al., 2012) although it does not harbor any predicted aggregation prone regions. It was also showed that AT3$_{182–291}$does not decrease the thermodynamic stability of the folded domain and directly interact with the monomeric JD (Santambrogio et al., 2010, Masino et al., 2004, Chow et al., 2004). To investigate the possible role of AT3$_{182–291}$ in the formation of intermolecular interactions during the AT3 aggregation process, we performed aggregation kinetic experiments by means of ThT fluorescence assays. First, our results show that AT3$_{182–291}$ is not responsive to ThT assay, which provides an experimental proof that this fragment is not amyloidogenic on its own (Figure S1). Then, we co-incubated a pathological variant of AT3 (AT3-Q55) with AT3$_{182–291}$at different molar ratios (Fig. 1A). AT3$_{182–291}$ is able to diminish the aggregation kinetics of the expanded AT3-Q55 in a concentration-dependent manner. This effect is not induced by increasing the lag phase of the aggregation kinetics but rather the elongation phase, suggesting that AT3$_{182–291}$ somehow competes with the full protein in the formation of oligomers thus slowing their formation process. Furthermore, co-incubation of AT3$_{182–291}$ with an AT3 variant truncated before the polyQ (AT3-291Δ) resulted in a similar effect (Fig. 1B). Therefore, it is likely that AT3$_{182–291}$ is able to interact with the protein region upstream of the polyQ stretch during the aggregation process. These data fit well with a previous study showing a protective effect against aggregation of polyQ proteins by AT3 UIMs (Miller et al., 2007). It is evident that AT3$_{182–291}$ plays a key role in the aggregation process of AT3 in that: i) it increases the aggregation rate of JD (AT3-291Δ) (Santambrogio et al., 2012) ii) it is probably involved in the intermolecular interactions formed during the AT3 aggregation process and iii) its presence (without the polyQ) is enough to trigger toxicity in a mouse model (Hubener et al., 2011).

Nevertheless, information at the atomic level on this crucial region of ataxin-3 is still missing, with the exception of a recent study in which the conformation of only its UIMs region in complex with Ub was investigated (Song et al., 2010). In light of the above observation, we provide here the first characterization of the structural properties of the AT3 disordered region by biophysical techniques and explicit solvent atomistic simulations.

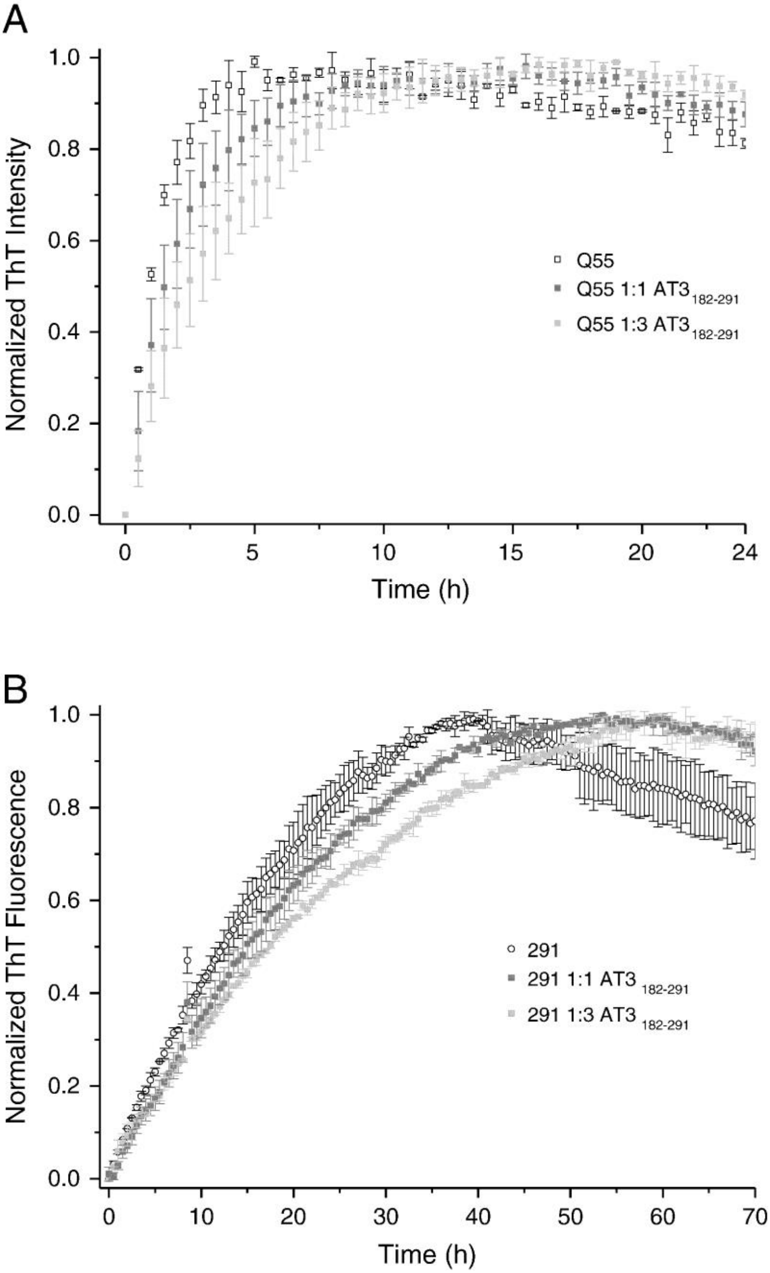


Figure 1. Effect of AT3$_{182-291}$ on the aggregation of AT3-Q55 and AT3-291Δ. Tht fluorescence were recorded at different concentration of AT3$_{182-291}$ co-incubated at 37 °C with 12.5 µM AT3-Q55 (A) or AT3-291Δ (B). Error bars are relative to standard deviation calculated on at least three different experiments.

AT3$_{182–291}$ is classified as an intrinsically disordered domain

The phase diagram for intrinsically disordered proteins (IDPs) is known to indicate mean hydrophobicity (<H>) and mean net charge as discriminating order parameters (Uversky et al., 2000). The <H> value for AT3$_{182–291}$ is around 0.38 (Figure S2), placing it in the region of the natively unfolded proteins in the two-dimensional diagram developed by Uversky and coworkers (Uversky et al., 2000), whereas the JD (residues 1–182) is located in the natively folded region of the graph.

When the contribution of the 182–291 domain to the two known AT3 isoforms (Goto et al., 1997) or the AT3-291Δ is considered, a shift toward the interface between natively folded and unfolded proteins can be observed. This result is consistent with the notion from biophysics experiments, which point out a mainly disordered structure of the AT3 C-terminal region (Maino et al., 2003) and allow us to ascribe AT3$_{182–291}$ to the group of ID proteins and domains.

Structural and dynamical characterization of the ID domain 182–291 of AT3

We integrated multiple independent atomistic molecular dynamics (MD) simulations (collecting more than 500 ns of MD ensemble overall) to experimental biophysical characterization of the domain by different techniques (CD, SEC and ESI-MS). The simulations were carried out starting from a structural model of AT3$_{182–291}$ with no side-chain side-chain intramolecular interactions (see Materials and methods). In fact, the model selected as a starting structure for the MD simulations features a 4.41 nm Radius of gyration (Rg), fully consistent with a structure lacking any artificial long-range (distance in sequence higher than 5 residues) intramolecular interactions (Figure S3) and a predicted α-helical content of 70% which includes the α-helices UIMs solved by NMR (Song et al., 2010). Particular attention was devoted to validating the MD ensemble by comparison with the experimental data. In particular, we evaluated the congruence between the Rg estimated by SEC and the Rg achieved in the simulations, as well as between average secondary structure content derived by CD and the secondary structure content from the simulations. α-Helix propensity was shown to be a hallmark of several IDPs structural ensembles (Dyson and Wright, 2005, Kjaergaard et al., 2010, Belle et al., 2008). AT3 harbors two UIMs (residues 224–243 and 244–263) in the 182–291 region, which are essential for its physiological function. The structures of UIM motifs from several other proteins are experimentally known (Fisher et al., 2003, Wang et al., 2005). They are generally α-helical in solution both in the free and Ub-bound state. The Ub-bound conformation of AT3 UIMs (AT3-UIM1-2: residues 222–263, PDB ID: 2KLZ) was recently solved by NMR spectroscopy (Song et al., 2010), which showed an α-helical structure with a typical compact helix–loop–helix fold. Nevertheless, atomistic details of the whole AT3$_{182–291}$ unbound ensemble are still missing. Thus, we employed CD spectroscopy to estimate the secondary structure content of the AT3$_{182–291}$ variant (Fig. 2A). Spectra deconvolution highlighted a highly disordered structure (63.7% random) with 22.4% of α-helices, in good agreement with the average secondary structure content from the MD ensemble (Fig. 2B) (26% α-helices and 60% random-coil, see below for details). The analysis of the MD ensemble also allowed the identification of the residues that are likely to adopt α-helical structures in solution in the free state of AT3$_{182–291}$. α-helical structures were estimated by DSSP secondary structure definition, which is based on the pattern of main-chain H-bonds, and by the g_helix Gromacs tool that accounts for both the H-bonds pattern and the psi and phi angles. The analysis shows that after only 50 ns of simulation, the residual helical content is 26% down from the initial 70% of the starting model whereas the random coil content rises up to 65% (Fig. 2C). These values are in good agreement with those experimentally obtained by CD spectroscopy experiments (22.4% α-helices and 63.7% random coils). Moreover, a residue-dependent profile of secondary structure persistence was calculated for the macro-trajectory from the DSSP definition (Fig. 2D and Table S1). This profile suggests five main regions characterized by α-helical structures with high persistence values: residues 187–202 (region I), residues 208–216 (region II), residues 227–233 (region III), residues 245–253 (region IV), residues 279–288 (region V). The analysis was also integrated by a residue-dependent profile derived by the g_helix Gromacs tool that reports the percentage of frames in which the residue assumes a helical conformation (Fig. 2E). The g_helix profile confirms the presence of the aforementioned five α-helical regions (I–V). Regions II and V are significantly stable in the MD ensemble (~ 70% of frames), whereas the three other regions (I, III and IV) feature a lower persistence in helical states (~ 50% of frames). The DSSP residue-dependent profile also suggests that regions II (52–99.1%), III (36.7–64.5%) and V (61.9–99.8%) mainly populate α-helices (Fig. 2 and Table S1). Indeed, the regions I and IV are most frequently found in π-helical structures with persistence values in the range from 43.5% to 88.8% (Fig. 2 and Table S1). π-helices are known to be unstable structural elements and they are generally shorter than 10 residues (Hollingsworth et al., 2009). In line with this observation, the π-helices in the AT3$_{182–291}$ MD ensemble are highly dynamical as also attested by the high standard deviation associated with them (around 4%). Nevertheless, the high occurrence of π-helices could also be due to artifacts intrinsic in the Gromos96 43a1 force field used in the simulations (Daura et al., 1998). The two analyses can provide a complementary view on the secondary structure propensity in the MD ensemble, since they are based on different geometrical parameters to define the secondary structures.

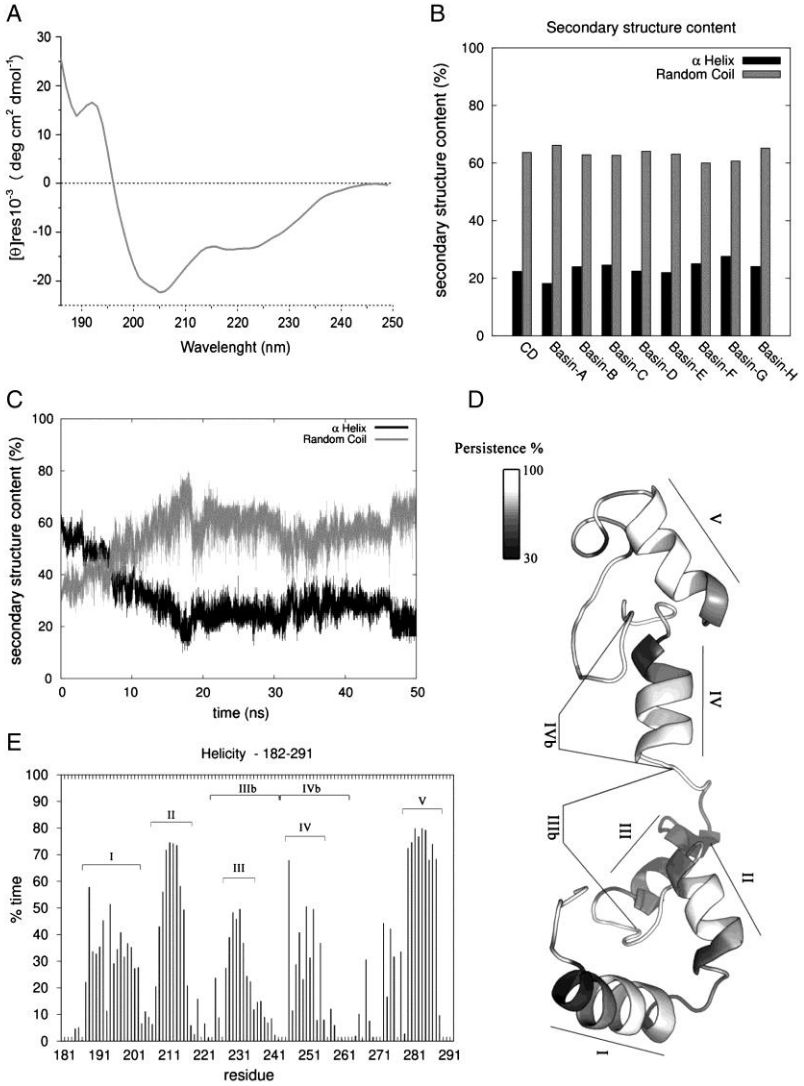


Figure 2. Secondary structure characterization. A. Circular dichroism spectrum in the near-UV of 10 μM AT3$_{182\text{-}291}$. B. Secondary structure content of AT3$_{182\text{-}291}$ obtained from CD experiments in comparison with values from MD ensemble. The secondary structure content was calculated both from structure obtained with cluster analysis and as average on each FEL basin structures, and expressed as percentage over the total number of residues in the domain. C. Evolution of secondary structure content during the first 50 ns of one MD trajectory as an example, starting from the extended model. D. Persistence of secondary structure calculated over the macro-trajectory and represented by a color gradient, from black (30 % of persistence) to white (100 % of persistence). The identified regions characterized by helical elements are highlighted (I-V), as well as the regions encompassing the two UIMs (IIIb, IVb) E. Residue-dependent profile of helicity. The identified regions characterized by helical elements are highlighted (I-V), as well as the regions encompassing the two UIMs (IIIb, IVb).

The two UIMs of the $AT3_{182–291}$ region consist of about 20 residues each: one from residue 224 to 243 (region IIIb) and one from residue 244 to 263 (region IVb) (Fig. 2D, E). The helical portions of the UIMs in free $AT3_{182–291}$ are generally shorter than in the Ub-bound state (Song et al., 2010) (Fig. 2D, E) and their C- and N-terminal extremities can populate different secondary structures from helices, π-helices or even random-coil structures in the MD ensemble. As stated above, UIM motifs in proteins have an intrinsic propensity to adopt helical conformations. Moreover, a well-known hallmark of IDPs is the formation of transient secondary structure elements upon binding to a partner (Dyson and Wright, 2005). Indeed, the UIMs in the NMR Ub-bound conformation (Song et al., 2010) showed a higher α-helical content (26%) with respect to that estimated by our simulation over the total length of the fragment. Therefore, α-helices are suggested to occur or to be stabilized in $AT3_{182–291}$ upon Ub-binding.

AT3 $_{182–291}$ domain populates both extended and compact states

To experimentally determine the Rg of $AT3_{182–291}$ size exclusion chromatography (SEC) was performed using proteins with a known 3D structure (carbonic anhydrase II, bovine β-lactoglobulin, bovine aprotinin, bovine ribonuclease A and bovine cytochrome C) as calibration standards. The Rg values of these proteins were determined by the Gromacs g_gyrate tool. The resulting distribution coefficients (Kd) plotted against Rg values were fitted with a linear function (correlation coefficient ≈ 0.99) (Fig. 3A). Two peaks for $AT3_{182–291}$ were identified with Rg values of 1.95 ± 0.074 nm (Stoke radius 2.44 ± 0.095 nm) and 2.26 ± 0.003 nm (Stokes radius 2.84 ± 0.004 nm), respectively (Fig. 3B). These values could be unambiguously assigned to different conformations of fragment monomer as confirmed by ESI-MS mass deconvolution analysis (see below for comments and see Fig. 3C), as well as native protein electrophoresis on native and denatured samples (Figure S4). Theoretical Stokes radii calculated for an IDP with the same molecular mass as $AT3_{182–291}$ in a coil and pre-molten globule states are 3.0 ± 0.32 nm and 2.6 ± 0.45 nm, respectively (Uversky, 2012). These values suggest that AT3 $_{182–291}$ exists in solution as a pre-molten globule with at least two different degrees of compactness.

The presence of two different main populations in solution was also suggested by nano-ESI-MS experiments executed by Dr. G. Invernizzi in the group of Prof. P. Tortora at the University of Milano-Bicocca. ESI-MS is suitable to identify different conformations of a protein in solution by means of charge state distributions (CSDs) (Kaltashov and Abzalimov, 2008). Its ability to preserve weak interactions proves to be particularly useful (van den Heuvel et al., 2004) and it has been recently applied to the study of conformational heterogeneity in IDPs (Woods et al., 2011). Charge states are dependent on the solvent accessibility of the protein surface and they are related to the protein compactness in solution. ESI-MS spectra of the isolated $AT3_{182–291}$ fragment clearly resulted in a bimodal distribution, with CSD centered on 14 + and 8 + peaks for the less and more compact conformations respectively (Fig. 3C). Interestingly, the same bimodal behavior was previously observed in the context of the whole AT3-291Δ variant, in which both extended and globular states related to $AT3_{182–291}$conformations were detected (Santambrogio et al., 2012). Integrated to our results, these observations, point out an intrinsic property of the domain to populate different conformational states. Although it could be tempting to extrapolate quantitative results from ESI-MS data, it should be noted that electrospray ionizations of proteins require solution with very low ionic strengths. Therefore, a direct comparison to our other data cannot be performed, since they were collected at a physiological ionic strength. Furthermore, like many IDPs (Uversky et al., 2000), $AT3_{182–291}$ is rich in charged residues, which might be responsible for tertiary contact (see section on Intra-molecular interactions) and are influenced by the ionic strength. The common properties of the isolated $AT3_{182–291}$ and of the 182–291 region in the context of the whole AT3 protein, contribute to enforcing the notion that identifying the conformational ensemble for the isolated domain is the first mandatory step in the attempt to define the structural characteristics of the whole AT3-291Δ variant.

To validate the description of the conformational ensemble achieved by MD simulations of $AT3_{182–291}$, the distribution of the Rg values was calculated over the macro-trajectory and compared with the experimental data obtained by SEC (Fig. 3D and S5). Notably, it turned out that two major populations were present with Rgs of 1.93 and 2.11 nm, respectively. These values are in very good agreement with the experimental ones (1.95 and 2.26 nm). Other less frequent structural states in the conformational ensemble can also be identified around 1.5 nm, 1.8 nm and 2.4/2.3 nm. It should be noted that the most compact state at 1.5 nm is unlikely to be representative of the protein in solution, since it would have been detected by SEC analysis. Furthermore, this Rg value is equal to a Stoke radius of 1.91 nm (assuming a globular conformation of the protein), which would fit to a value corresponding to a globular folded protein of the same mass of $AT3_{182–291}$(Uversky, 2012). Over-compaction of unfolded proteins is a common problem for current MD force fields and has been shown for other IDPs (Knott and Best, 2012). This strongly encourages the integration of MD data with biophysical techniques to characterize their conformational ensemble. Interestingly, a 200 ns simulation with the newly released force field, CHARMM22star (Piana et al., 2011), which was also proven promising for simulating unfolded proteins (Lindorff-Larsen et al., 2012), does not feature populations with Rg values lower than 1.95 nm. This enforces the notion that the over-compact states with Rg values lower than 1.5 nm are not meaningful (*data not shown*). Notably, some of the residues located in the two UIMs tend to lose their α-helical conformation (as previously shown, Fig. 2) in populations with a less compact structure while they re-assumed an α-helical structure (*data not shown*) in population with a more compact conformation.

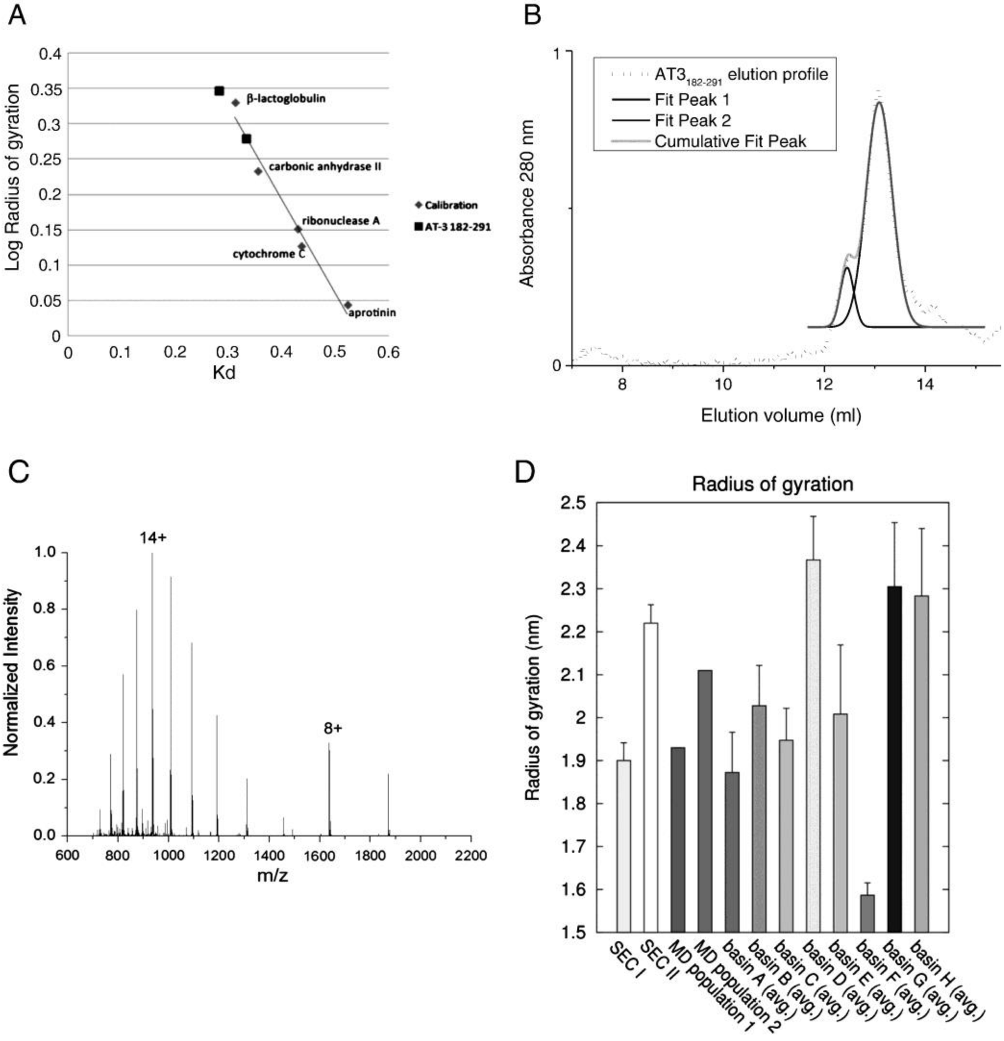


Figure 3. AT3$_{182-291}$ conformational states in solution. A. SEC profile of AT3$_{182-291}$ was multi-fitted with standard Gaussian equations (R2 =0.96) and elution volumes determined as y0 of the equation. B Kd values of AT3$_{182-291}$ (red squares) plotted against a reference set of standard proteins with known Rg values (blue diamonds). Coordinates of reference proteins were fitted to a linear equation (y = -1.3259x + 0.7226) and Rg values for AT3$_{182-291}$ calculated. C. ESI-MS spectrum of 5 µM AT3$_{182-291}$ shows charge state distribution centered at 8+ and 14+ z values representative of compact and less compact conformations respectively. D. The Rg values identified from SEC experiments (indicated as SECI and SECII) are compared to those obtained from fitting the distribution of the Rg values calculated over the macro-trajectory (the two major populations are indicated as MD population 1 and MD population 2) and each conformational basin (basin A-H).

It is also worth mentioning that among the compact conformations sampled in the MD ensemble, a structure similar to the UIMs Ub-bound conformations (PDB ID: 2KLZ, (Song et al., 2010)) was not identified. In fact, in the Ub-bound states the two UIMs interact by intramolecular hydrophobic contacts and assume an α-helix conformation for their entire length. The comparison between structures from the MD ensemble and the structure of the Ub-bound UIMs was carried out by both monitoring distances between specific pairs of residues, such as L233–I253, and by a different set of rmsd calculations. Rmsd calculations were carried out with a different subset of mainchain atoms that accounts for different residues within the UIM stretches. Compact structures, in which UIMs assume an α-helical content similar to the Ub-bound state, can be identified (*data not shown*). Nevertheless, at the tertiary level, the MD ensemble does not contain AT3$_{182–291}$ conformations similar to the bound UIMs. The rmsd between MD frames and Ub-bound UIMs ranged from 0.5 to 6 nm. The distance between L233 and I253 was very large (about 1.25–1.5 nm), whereas in the NMR bound structure they were about 0.6 nm apart. Despite the fact that we did not detect bound-like conformations in the free ensemble of AT3$_{182–291}$, this domain may still sample them in the free states. In fact, the lack of detection of bound-like conformations could either be due to intrinsic limits in the sampling achieved by our simulations, or to the fact that we are comparing a NMR structure of only the two UIMs portion (residues 222–263) with that of the whole disordered fragment (182–291). Our results therefore suggest that AT3$_{182–291}$ can populate minor compact states in which the UIMs feature an α-helical content compatible with the known Ub-bound conformations. Despite this similar arrangement, the free UIM conformations in these minor states are still remarkably different from the bound states. Conformational changes, likely to be promoted or enhanced by the interaction with Ub, are therefore required to bring the hydrophobic residues at a distance lower than 0.6 nm.

AT3$_{182–291}$ structural ensemble in solution

Considering the high heterogeneity of the IDP ensemble, structural clustering based on mainchain rmsd matrices can be difficult and it cannot account for the different substates populated by the protein in solution. Therefore, to achieve a higher resolution description of the conformational landscape of AT3$_{183–291}$, Cα and all-atom PCA were carried out on the macro-trajectory.

The projection of the MD trajectories on the first two eigenvectors calculated by PCA show both the capability of the system to explore different basins of the densely populated subspace, and an overlap between conformations obtained from independent simulations (Figure S6). In fact, the rmsip of the first 10 PCs (accounting for more than 89% of the total variance in our trajectories, Figure S7A) of the independent MD replicates is a quantitative similarity index between subspaces sampled by different trajectories in the context of the essential subspace captured by the PCA (Amadei et al., 1999). In our

MD simulations, rmsip ranges from 0.67 to 0.77, showing that a sufficient overlap was achieved in the simulations, (Figure S7B). The order parameter *O* was calculated to better quantify the heterogeneity of the conformational ensemble described by the MD simulations of AT3$_{182–291}$ (Fisher and Stultz, 2011). Indeed, this parameter can be considered as a quantitative measurement of the disorder in the structural ensemble under investigation. The order parameter was calculated on the average structures derived from cluster analysis on the MD ensemble (see Materials and methods for details). The results are reported in Table 1. The AT3$_{182–291}$ domain displays very low *O* parameter values, ranging from 0.178 to 0.199 depending on the number of clusters considered. Such values agree with the ones previously reported for other IDPs (Fisher and Stultz, 2011). The limit value of 0 applies to the ideal case of an infinite number of equally populated, different conformations. Therefore, these results overall indicate a highly heterogeneous conformational ensemble for AT3$_{182–291}$. To achieve a two-dimensional representation of the conformational landscape described by our MD ensemble, an estimation of the Free Energy Landscape (FEL) was carried out (Fig. 4A) at first using the projections of the MD macro-trajectory along the first and the second PCs as reaction coordinates. These projections account for more than 50% of the total variance (Figure S7A). Although not accurate enough to exactly calculate free energy barriers between the basins, this FEL representation allows a description of the conformational landscape accessible to the molecule in the simulations and to disclose different conformational states (Zhuravlev et al., 2009). Three main basins were identified on the FEL (namely A, B and C), along with other regions (F, E, D, G and H) (Fig. 4A) characterized by a lower occurrence probability and therefore associated to higher energy values. Each MD simulation transiently samples different basins (Figure S6). This confirms that the simulations are not entrapped in local minima and supports the notion that the AT3$_{182–291}$ fragment in solution can populate different conformational states. The structures populating each of the different basins on the FEL were isolated and a structural cluster analysis was carried out on each of them to define a reference average structure, which features the lowest mainchain rmsd with respect to all the other structures in the same basin. The distribution of the Rg values along the FEL is reported in Fig. 4B. As discussed above, it turns out that there are different sub-populations with different degree of compactness that are ascribable to the pre-molten globule class (Uversky, 2012). Noticeably, basins A, B and C, which are the most populated basins in the MD ensemble, have average Rg values of 1.87, 2.03 and 1.95 nm respectively (Table S2).

Moreover, according to PCA the AT3$_{182–291}$ can be divided into two dynamical sub-domains; the N-terminal (residue R182-D241; blue, Fig. 4A) and the C-terminal (residue M242-K291; green, Fig. 4A) subdomains. Different conformational states in the FEL are related to different reciprocal orientations of the two sub-domains, as detailed in the next section. In particular, there are more compact states (as in basins A and C) in which the two sub-domains interact with each other, and less compact states (as in basin B), which feature intra-subdomain interactions but not inter-subdomain interactions.

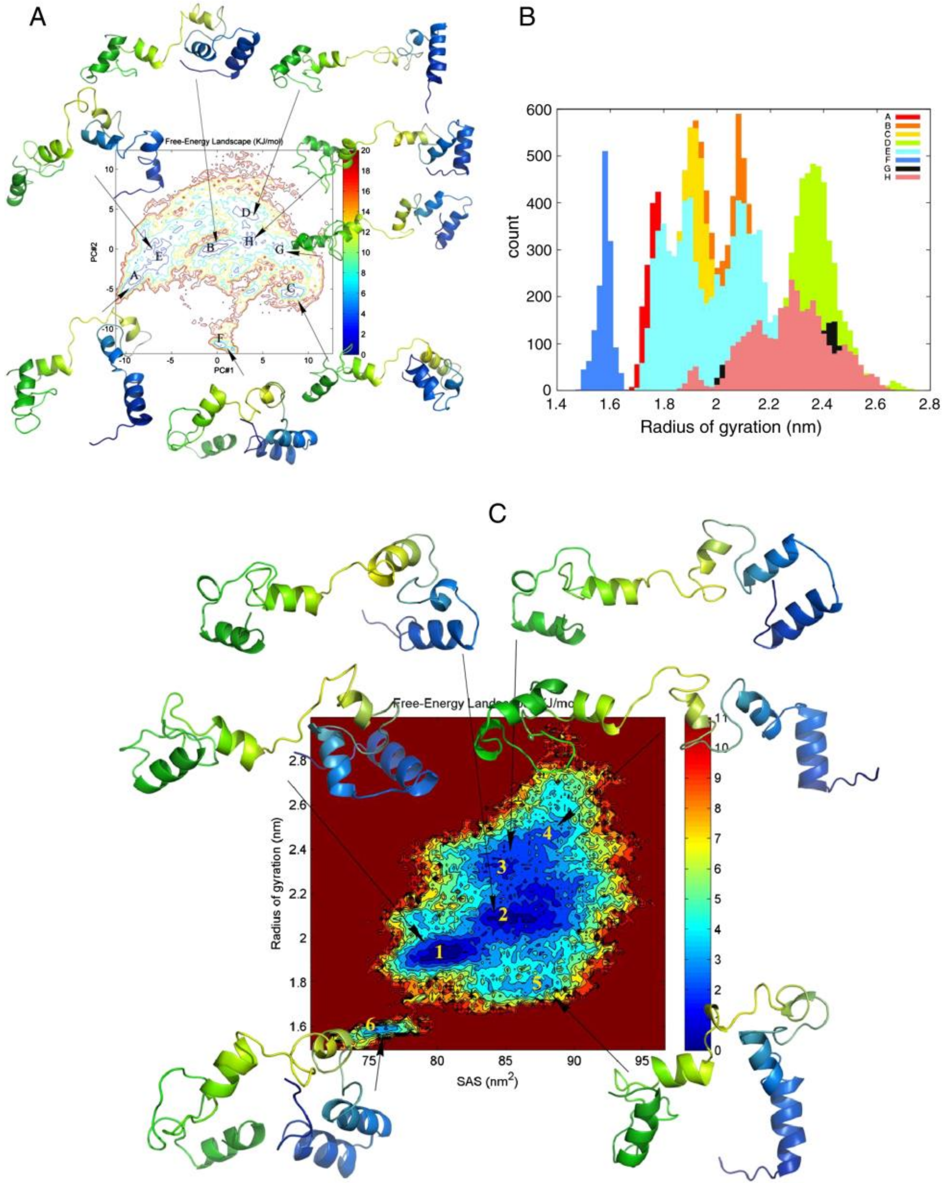


Figure 4. FEL representation of AT3$_{182\text{-}291}$. A. FEL representation using the projection of the macro-trajectory along the 1st and 2nd principal components as reaction coordinates. Free energy is given in kJ/mol as indicated by the color bar. The major basins are labeled by capital letters (A to H). The average structures calculated for each conformational basin are represented as cartoons and highlighted by a color gradient, from N-terminus (blue) to C-terminus (green). B. Distribution profiles of the Rg

values for each conformational basin shown in panel A. C. FEL representation calculated using the SAS values and Rg values as reaction coordinates. The free energy is given in kJ/mol and indicated by the color code shown in the figure. The major basins are labeled by numbers (1 to 6). The average structures calculated for each conformational basin are represented as cartoons and highlighted by a color gradient, from N-terminus (blue) to C-terminus (green).

To better describe the different states with respect to Rg values, a FEL representation using the Rg and the Solvent Accessible Surface area (SAS) as reaction coordinates was carried out (Fig. 4C). This FEL points out the presence of two main sub-populations, basin 1 and 2 respectively that are characterized by different degrees of compactness in terms of Rg values. The first main basin is characterized by SAS values around 80 nm^2 and by an average Rg value of 1.93 nm (Fig. 4C, basin 1), corresponding to the more compact conformations, whereas the second one has SAS values around 85 nm^2 and an average Rg of 2.11 nm (Fig. 4C, basin 2), corresponding to the less compact states. These two main basins also fit the experimental Rg values estimated by SEC. Furthermore, the structures that populate the four other regions on the FEL were isolated and investigated in more detail (basin 3, 4, 5, 6) (Fig. 4C and Table S3). The basins 3 and 4 comprise structures that have an even lower degree of compactness, showing average Rg values of 2.31 nm and 2.51 nm and average SAS values of 85.35 nm^2 and 88.70 nm^2 respectively (Table S3). Indeed the basins 5 and 6 comprise more compact structures, with average Rg values of 1.79 nm and 1.59 nm and average SAS values of 86.36 nm^2 and 76.01 nm^2 respectively (Table S3). Rg values comparable to the ones observed in basins 5 and 6 were not detected by experimental analysis, suggesting that they are likely to be artifacts, caused by the bias for over-compaction of unfolded proteins in the commonly employed force fields for MD (Knott and Best, 2012), as also discussed above. These two basins were therefore not considered for further analyses. Their appearance in the MD ensemble also strengthens the need to integrate experimental biophysical spectroscopies with MD simulations of IDPs.

Intra-molecular interactions

In the attempt to identify the molecular determinants of the compact states of AT3$_{182–291}$, the different classes of intra-molecular interactions were calculated in the MD *ensemble* along with their persistence (Table S4–S7). We carried out the analysis by *PyInteraph* (Tiberti et al. 2014). We then estimated the threshold of persistence that was 20% to discard non-significant and poorly populated interactions in the ensemble. The networks of interaction were analyzed by *xPyder* tools (Pasi et al. 2013). The results show that AT3$_{182-291}$ has large networks of salt-bridges, composed by a high number of transient interactions. It turned out that the more compact states have a higher number of salt bridges interactions and more interconnected networks compared to less compact states (Figure 5). Moreover, the analyses identified a higher number of hub residues (i.e. residues connected to more than 3 nodes in the graph) in the more compact states (Figure 6) that can have a key role in the formation of tertiary structures. Aromatic-aromatic interactions are absent and only a few amino-aromatic interactions could be detected. We can find Y288 among the residues involved in amino-aromatic interactions with highest persistence. It plays a central role in all the FEL basins (Figure S8). Despite the occasional high persistence, amino-aromatic interactions seem to have only a marginal role in the regulation of tertiary contacts in AT3$_{182–291}$compared with the multiple and very persistent salt-bridges. Hydrophobic interaction networks are also present (Figure S9), but they are generally characterized by very low persistence. A high number of transient hydrogen bonds are present in the entire domain (Figure S10). Hydrogen bonds with higher persistence (more than 70%) are associated with helical elements. As other IDPs (Mao et al., 2010), AT3$_{182–291}$ has a very high content of charged residues (42 charged residues out of 109), which are interconnected to form electrostatic networks. Indeed, it was recently proposed that these networks may play a crucial role in the definition of IDP properties, as disorder/order transitions, promotion or maintenance of globular state (Zhang et al., 2012). The results show that AT3$_{182–291}$ is characterized by the presence of a high number of electrostatic interactions, which form large networks extended throughout the protein (Fig. 5). To characterize the electrostatic interactions and their networks, we employed the xPyder plugin (Pasi et al., 2012), a Pymol plugin that carries out analyses inspired by graph theory for the study of intramolecular interactions. To reduce the number of non-significant interactions, all the salt-bridges of a basin with persistence higher than 20% were considered for the analysis. Each residue belonging to a salt-bridge was represented as a node of an undirected and unrooted graph and edges were built between the nodes according to the persistence of the interaction in the MD ensemble. This representation allows us to analyse the topological properties of the network overall, and to propose putative paths of long-range communication between distal residues.

The conformations in the MD *ensemble* show that (Fig. 5) compact states identified in basin 1 (average Rg value of 1.93 nm, indicated as 1 in Fig. 5), have a higher number of pair-wise salt bridges (30 and 24 respectively, Table S4) compared to less compact states, in basin 2 (average Rg value of 2.11 nm, indicated as 2 in Fig. 5). Basin 1 conformations also show salt bridge networks composed by highly interconnected residues (Table S4 and Fig. 5). This is also true when basin 1 structures are compared to other less compact states, as the ones sampled in basins 3 and 4 (Fig. 4C and Table S4). These results suggest that the salt bridges are a major determinant for the structural properties of AT3$_{182–291}$. In this context, individual salt-bridges have usually a transient nature, so each conformational state is stabilized by different pairs of interacting residues, determining the formation of wide but also highly plastic interaction networks.

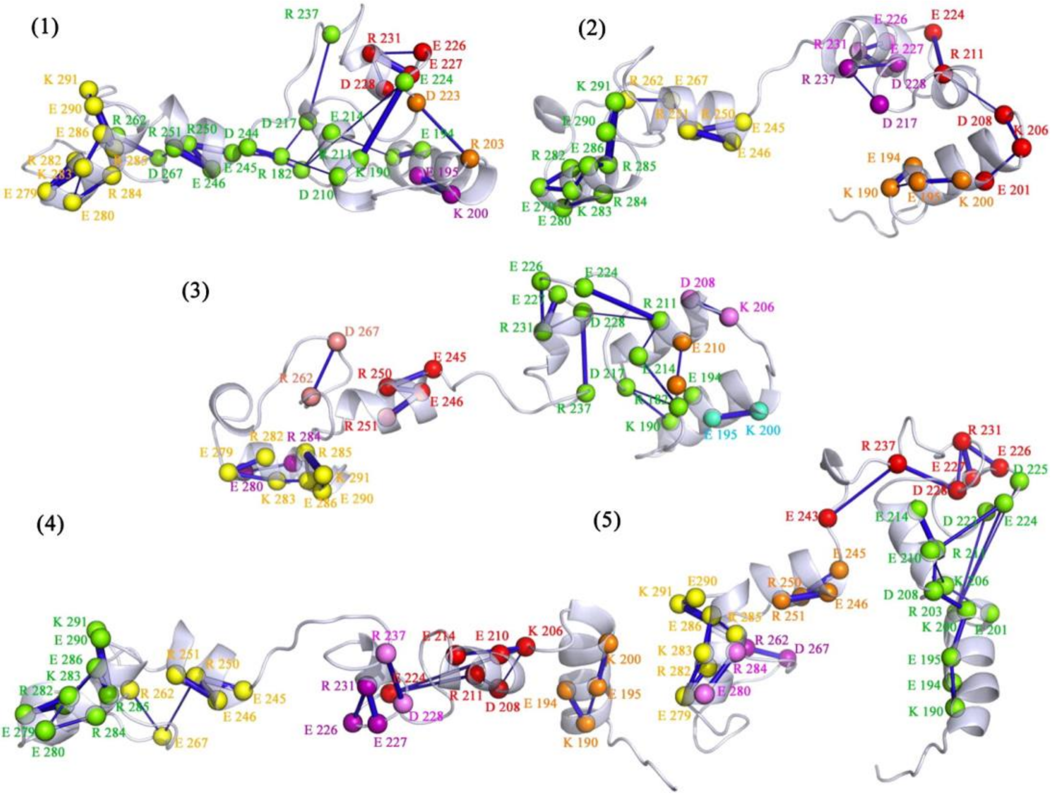


Figure 5. Networks of salt bridge interactions of AT3$_{182\text{-}291}$. The average structure of each conformational basin of the FEL (from 1 to 5) is shown as cartoon, and the Cα atoms of the acidic and basic residues involved in salt bridges are shown as spheres, respectively. The Cα atoms of the interacting residues are connected by sticks whose thickness is proportional to the persistence of the interaction. The single sub-networks are represented with a color code indicating their respective size: green-yellow-orange-red-purple-magenta-cyan-pink sort out from the highest size to lowest size. A detailed list of the salt bridges is reported in Table S4.

Hub residues, which are defined as highly interconnected residues in protein structure networks (i.e. residues connected with more than 3 neighbors), are crucial residues for the maintenance of a protein fold and therefore putative hot spots for tertiary structure formation (EsAngelova et al., 2011, Vishveshwara et al., 2009). AT3$_{182–291}$ shows a higher number of hubs (4 vs. 2) in the most populated and more compact states (i.e. basin 1 structures). In particular, basin 1 and 2 structures have as common hubs R250 and K283 located in the C-terminal region, whereas in the basin 1 states also R182 and R231 act as hub residues (Fig. 6 and Table S8). The hub residues identified in AT3$_{182–291}$ are generally involved in interactions with several different partners (i.e. residues interacting with the hubs are not always conserved in the different structures) suggesting a major role for these amino-acids in the formation of electrostatic networks and an intrinsic plasticity of AT3$_{182–291}$.

Also, the sub-networks (also known as connected components) of salt bridges were identified (Fig. 5) by considering all the nodes that are linked by at least one path, but for which no path exists between the nodes of the connect component and the rest of the graph. Although the structures in the MD ensemble have a similar number of non-connected sub-networks (ranging from 5 to 8) these are generally composed by individual salt bridges or three/four-node paths in less compact structures, while compact states generally feature sub-networks with a higher number of interconnected residues (Fig. 5). The main sub-network (i.e. the sub-network accounting for a larger number of nodes) of each FEL basin was also analysed and they are reported as green spheres in Fig. 5. The more compact states account for the largest main sub-network, composed by 16 residues (R182–D267) that connect distant regions of the fragment.

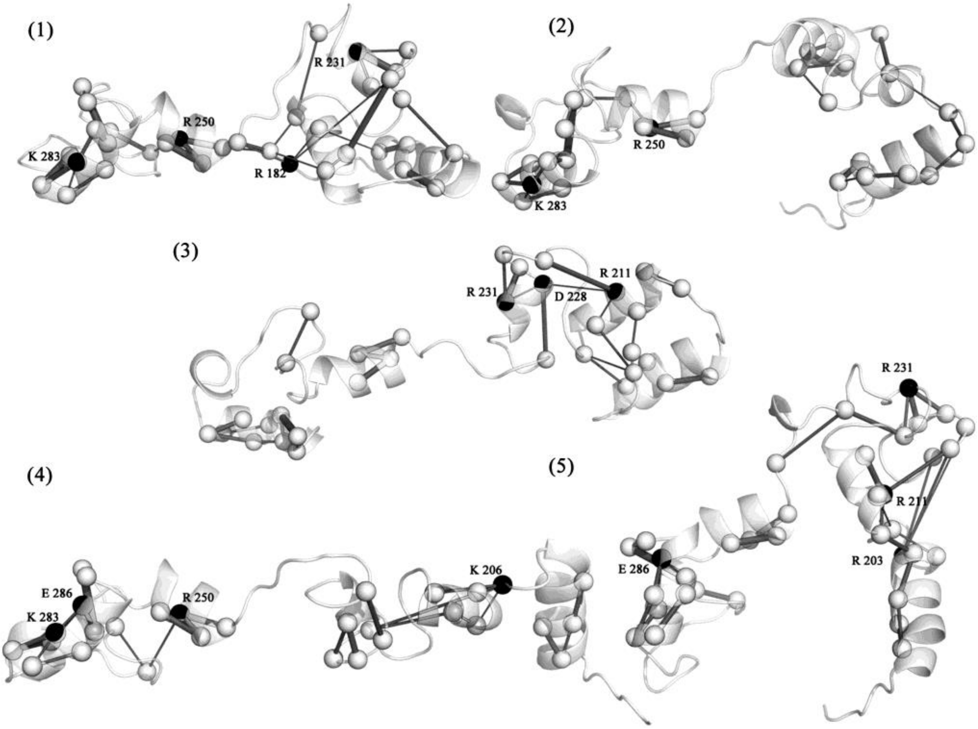


Figure 6. Hub residues in the salt bridge networks. The hub residues are defined as those involved in at least three different interactions. The salt bridges networks are plotted as cylinders on the average structure of each conformational basin of the FEL (from 1 to 5), shown as cartoon. The Cα atoms of residues involved in the interactions are shown as spheres, whereas Cα of the hub residues are highlighted in yellow. A detailed list of the hub residues is reported in Table S8.

On the contrary, only 9 residues localized in the C-terminal region of the domain are involved in the main sub-network of the less compact states (Fig. 5, residue E279–K291). Each sub-network comprises salt bridges between residues localized in specific tracts of $AT3_{182–291}$. These tracts correspond to the previously identified helical elements (region I to V, Fig. 2).

The importance of salt-bridge interactions in promoting the tertiary structure of $AT3_{182–291}$ is also supported by the investigation of $AT3_{182–291}$ by ESI-MS in presence of 1% formic acid (Fig. 7).

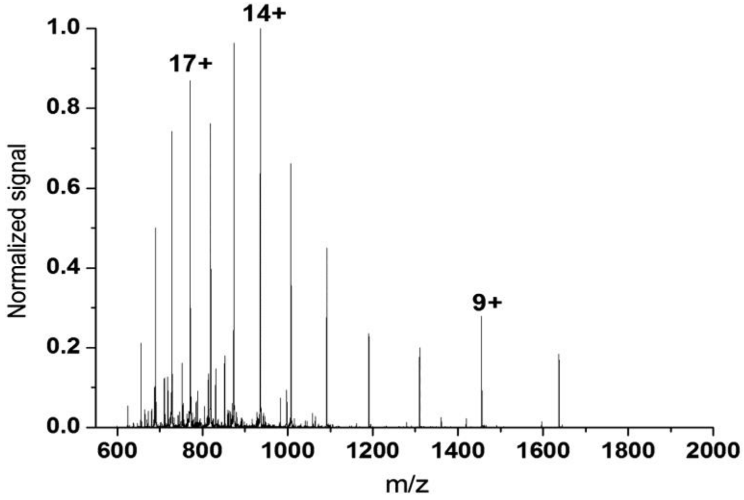


Figure 7. Effect of low pH on AT3182-291 structural conformations. ESI-MS spectra of 5 μM $AT3_{182-291}$ at pH 2.2. At least three different protein conformations are observable with charge state distribution centered at 9+, 14+ and 17+ z values.

In fact, at low pH (~ 2.2) the presence of CSDs was evident at higher charges (17 +) than those revealed at neutral pH. This result, together with less intense CSDs identifying the compact state (9 +), points to an increase in a more ionizable, hence extended conformation under experimental conditions unsuitable for salt bridge formation. Notably, the same degree of unfolding cannot be reached at neutral pH by increasing the experimental temperature, suggesting that compactness of $AT3_{182–291}$ is not affected by loss of hydrophobic or hydrogen bond interactions (Figure S11). Overall, our results show that charged residues and their networks of interactions can play a central role in modulating the dynamical and structural properties of $AT3_{182–291}$.

2.3.4 Concluding remarks

Polyglutamine proteins are generally complex multidomain polypeptides in which are present both globular folded and unstructured domains. In light of unveiling their physiological function, largely debated for the majority of them, and to understand the molecular mechanisms responsible for abnormal aggregation upon polyQ expansion, a clear description of their disordered regions is fundamental.

Moreover, the aggregation pathway and consequently the cytotoxicity of expanded polyQ proteins are known to be profoundly affected by polyQ flanking regions (Duennwald et al., 2006). In this context, the characterization of the domains flanking the polyQ is mandatory to understand the molecular process underlying their pathological aggregation. In the multidomain AT3, the N-terminal structured Josephin domain (JD) undergoes amyloid aggregation isolated in solution (Masino et al., 2004) and plays a key role in determining the early aggregated species of the whole protein (Ellisdon et al., 2006). Noteworthy, the disordered tract spanning from JD to polyQ greatly enhances the aggregation kinetics of the JD, even if it does not harbor any aggregation prone regions (APRs)(Santambrogio et al., 2012). Further, the AT3 variant truncated at residue 259, which is physiologically relevant as a cleavage product of caspases, was shown to be toxic and to induce an ataxic phenotype in a murine model (Hübener et al., 2011). In this light, the characterization of AT3 disordered tract is critical to understand the mechanism behind AT3 aggregation and toxicity, for which only few structural information are available to date.

Here, we have demonstrated that AT3$_{182\text{-}291}$ is a monomeric intrinsically disorder domain in solution. We have also experimentally demonstrated that the AT3$_{182\text{-}291}$ fragment does not feature aggregation propensity by itself. In agreement with previous observations on other UIMs (Miller et al., 2007), we also show that it can exert a protective effect against aggregation of the entire AT3 likely to be related to the formation of intermolecular interactions by the UIM motifs. In fact, UIMs motifs in AT3 were found to interact with each other upon binding to ubiquitin (Song et al., 2010). Interestingly also other UIMs, as the ones isolated from the Vps27p, were shown to mutually interact (Fisher et al., 2003). Moreover, preliminary protein-protein docking by Z-dock (Pierce et al., 2005) using the UIM-1 and UIM-2 structures isolated from the 2KLZ pdb entry (Song et al., 2010), showed top score complexes with propensity to form inter-UIMs interactions in AT3 compatible with the ones observed in the Ub-bound structure, thanks in particular to a complementarity between their acidic and basic residues (*data not shown*). Noteworthy, the AT3$_{182\text{-}291}$ UIM motifs are particularly rich in Arg residues. Arginine is a suitable residue to mediate networks of electrostatic interactions, allowing up to two salt-bridges and five H-bonds formation thanks to the unique properties of the guanidine group (Cupo et al., 1980; Mrabet et al., 1992).

The heterogeneous and dynamic nature of IDPs makes their structural characterization in the free state very challenging. MD force fields were proven useful to characterize the conformational landscape of IDPs and unfolded proteins (Espinoza-Fonseca, 2009; Knott and Best, 2012; Lindorff-Larsen et al., 2012; Arrigoni et al., 2012). Nevertheless, computational investigation of IDPs by MD simulations alone still suffers of limitations. The integration of MD simulations with experimental techniques can therefore improve the quality of the results achieved in the structural characterization of IDPs (Esteban-Martín et al., 2012).

The atomistic description provided in this work by the integration of different biophysical techniques and multiple all-atom explicit solvent MD simulations ascribes $AT3_{182-291}$ to the class of IDPs and highlights its propensity to sample different structural conformations with a different degree of compactness. Noteworthy, this disordered tract is able to assume two main sub-states, which can be ascribed to the pre-molten globule class and characterized by a high heterogeneity in terms of secondary and tertiary structures. Our experiments and the network description of the intramolecular interactions in the MD ensemble also point out a major role for salt-bridge interactions in the maintenance and/or promotion of the tertiary contacts observed for the fragment. A pivotal role is suggested for residues R211, R231, R250, K283, which acts as conserved hubs in the network and can be a suitable subset of residues for further investigations. The results here collected also provide an important framework to more detailed characterization of the fragment with NMR and SAXS spectroscopies, which have been proved very effective in capturing the complicate nature of IDPs (Salmon et al., 2012; Bernadó et al., 2007; Bernadó and Svergun, 2012; Kjaergaard and Poulsen, 2012).

2.3.5 Acknoledgments

This research was supported by computational facilities at Caspur HPC-Starting Grant 2011 and 2012 to EP, and GI. GI, PT and MER also gracefully acknowledge for funding Fondazione Cariplo (N.O.B.E.L. project), NEDD (Network Enabled Drug Design) of Regione Lombardia and the Italian Ministry for Instruction, University and Research (PRIN 2007: "A multidisciplinary approach to the study of *in vivo* and *vitro* aggregation of polyglutamine-containing proteins. Role of molecular and environmental factors" and project n. 2007XY59ZJ_004).

2.3.6 References

Amadei A., M.A. Ceruso, A. Di Nola, On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations, *Proteins*, 36 (1999) 419–424.

Angelova K., A. Felline, M. Lee, M. Patel, D. Puett, F. Fanelli, Conserved amino acids participate in the structure networks deputed to intramolecular communication in the lutropin receptor, *Cell Mol Life Sci*, 68 (2011) 1227–39.

Arrigoni A., B. Grillo, A. Vitriolo, L. De Gioia, E. Papaleo, C-Terminal acidic domain of ubiquitin-conjugating enzymes: a multi-functional conserved intrinsically disordered domain in family 3 of E2 enzymes, *J Struct Biol*, 178 (2012) 245–59.

Belle V., S. Rouger, S. Costanzo, E. Liquière, J. Strancar, B. Guigliarelli, A. Fournel, S. Longhi, Mapping alpha-helical induced folding within the intrinsically disordered C-terminal domain of the measles virus nucleoprotein by site-directed spin-labeling EPR spectroscopy, *Proteins*, 73 (2008) 973–88.

Bernadó P., E. Mylonas, M. V Petoukhov, M. Blackledge, D.I. Svergun, Structural characterization of flexible proteins using small-angle X-ray scattering, *J Am Chem Soc*, 129 (2007) 5656–64.

Bernadó P., D.I. Svergun, Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering, *Mol Biosyst*, 8 (2012) 151–67.

Chow M.K.M., A.M., Ellisdon, L.D., Cabrita, S.P. Bottomley,Polyglutammine expansion in ataxin-3 does not affect prtein stability: implications for misfolding and disease. *J Biol Chem* 279 (2004) 47643-47651.

Cupo P., W. El-Deiry, P.L. Whitney, W.M. Awad, Stabilization of proteins by guanidination, *J Biol Chem*, 255 (1980) 10828–33.

Duennwald M.L., S. Jagadish, P.J. Muchowski, S. Lindquist, Flanking sequences profoundly alter polyglutamine toxicity in yeast, *Proc Natl Acad Sci U S A*, 103 (2006) 11045–11050.

Dunker A.K., Z. Obradovic, The protein trinity--linking function and disorder, *Nat Biotechnol*, 19 (2001) 805–6.

Dyson H.J., P.E. Wright, Intrinsically unstructured proteins and their functions, *Nat Rev Mol Cell Biol*, 6 (2005) 197–208.

Espinoza-Fonseca L.M., D. Kast, D.D. Thomas, Thermodynamic and structural basis of phosphorylation-induced disorder-to-order transition in the regulatory light chain of smooth muscle myosin, *J Am Chem Soc*, 130 (2008) 12208–9.

Espinoza-Fonseca L.M., Leucine-rich hydrophobic clusters promote folding of the N-terminus of the intrinsically disordered transactivation domain of p53, *FEBS Lett*, 583 (2009) 556–60.

Esteban-Martín S., R. Bryn Fenwick, X. Salvatella, Synergistic use of NMR and MD simulations to study the structural heterogeneity of proteins, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2 (2012) 466–478.

Ellisdon A.M., B. Thomas, S.P. Bottomley, The two-stage pathway of ataxin-3 fibrillogenesis involves a polyglutamine-independent step, *J Biol Chem*, 281 (2006) 16888–16896.

Fisher R.D., B. Wang, S.L. Alam, D.S. Higginson, H. Robinson, W.I. Sundquist, C.P. Hill, Structure and ubiquitin binding of the ubiquitin-interacting motif, *J Biol Chem*, 278 (2003) 28976–84.

Fisher C.K., C.M. Stultz, Constructing ensembles for intrinsically disordered proteins, *Curr Opin Struct Biol*, 21 (2011) 426–31.

Goto J., M. Watanabe, Y. Ichikawa, S.B. Yee, N. Ihara, K. Endo, S. Igarashi, Y. Takiyama, C. Gaspar, P. Maciel, S. Tsuji, G.A. Rouleau, I. Kanazawa, Machado-Joseph disease gene products carrying different carboxyl termini, *Neurosci Res*, 28 (1997) 373–377.

Hollingsworth S.A., D.S. Berkholz, P.A. Karplus, On the occurrence of linear groups in proteins, *Protein Sci*, 18 (2009) 1321–5.

Hübener J., F. Vauti, C. Funke, H. Wolburg, Y. Ye, T. Schmidt, K. Wolburg-Buchholz, I. Schmitt, A. Gardyan, S. Driessen, H.-H. Arnold, H.P. Nguyen, O. Riess, S. Drießen, N-terminal ataxin-3 causes neurological symptoms with inclusions, endoplasmic reticulum stress and ribosomal dislocation, *Brain*, 134 (2011) 1925–42.

Kaltashov I.A., R.R. Abzalimov, Do ionic charges in ESI MS provide useful information on macromolecular structure?, *J Am Soc Mass Spectrom*, 19 (2008) 1239–46.

Keller B., X. Daura, W.F. van Gunsteren, Comparing geometric and kinetic cluster algorithms for molecular simulation data, *J Chem Phys*, 132 (2010) 074110.

Kjaergaard M., K. Teilum, F.M. Poulsen, Conformational selection in the molten globule state of the nuclear coactivator binding domain of CBP, *Proc Natl Acad Sci U S A*, 107 (2010) 12535–40.

Kjaergaard M., F.M. Poulsen, Disordered proteins studied by chemical shifts, *Prog Nucl Magn Reson Spectrosc*, 60 (2012) 42–51.

Knott M., R.B. Best, A preformed binding interface in the unbound ensemble of an intrinsically disordered protein: evidence from molecular simulations, *PLoS Comput Biol*, 8 (2012) e1002605.

Lindorff-Larsen K., N. Trbovic, P. Maragakis, S. Piana, D.E. Shaw, Structure and dynamics of an unfolded protein examined by molecular dynamics simulation, *J Am Chem Soc*, 134 (2012) 3787–91.

Mao Y., F. Senic-Matuglia, P.P. Di Fiore, S. Polo, M.E. Hodsdon, P. De Camilli, Deubiquitinating function of ataxin-3: insights from the solution structure of the Josephin domain, *Proc Natl Acad Sci U S A*, 102 (2005) 12700–12705.

Mao A.H., S.L. Crick, A. Vitalis, C.L. Chicoine, R. V Pappu, Net charge per residue modulates conformational ensembles of intrinsically disordered proteins, *Proc Natl Acad Sci U S A*, 107 (2010) 8183–8.

Masino L., V. Musi, R.P. Menon, P. Fusi, G. Kelly, T. a. Frenkiel, Y. Trottier, A. Pastore, Domain architecture of the polyglutamine protein ataxin-3: a globular domain followed by a flexible tail, *FEBS Letters*, 549 (2003) 21–25.

Masino L., G. Nicastro, R.P. Menon, F. Dal Piaz, L. Calder, A. Pastore, F.D. Piaz, Characterization of the structure and the amyloidogenic properties of the Josephin domain of the polyglutamine-containing protein ataxin-3, *J Mol Biol*, 344 (2004) 1021–1035.

Miller S.L.H., E.L. Scappini, J. O'Bryan, Ubiquitin-interacting motifs inhibit aggregation of polyQ-expanded huntingtin, *J Biol Chem*, 282 (2007) 10096–103.

Mrabet N.T., A. Van den Broeck, I. Van den brande, P. Stanssens, Y. Laroche, A.M. Lambeir, G. Matthijssens, J. Jenkins, M. Chiadmi, H. van Tilbeurgh, Arginine residues as stabilizing elements in proteins, *Biochemistry*, 31 (1992) 2239–53.

Nicastro G., S. V Todi, E. Karaca, A.M.J.J. Bonvin, H.L. Paulson, A. Pastore, Understanding the role of the Josephin domain in the PolyUb binding and cleavage properties of ataxin-3, *PLoS One*, 5 (2010) e12430.

Nozaki K., O. Onodera, H. Takano, S. Tsuji, Amino acid sequences flanking polyglutamine stretches influence their potential for aggregate formation, *Neuroreport*, 12 (2001) 3357–3364.

Pasi M., M. Tiberti, A. Arrigoni, E. Papaleo, xPyder: A PyMOL Plugin To Analyze Coupled Residues and Their Networks in Protein Structures, *J Chem Inf Model*, (2012).

Pasi, M., M. Tiberti, A. Arrigoni and E. Papaleo (2012). xPyder: a PyMOL plugin to analyze coupled residues and their networks in protein structures. *J Chem Inf Model* 52(7): 1865-1874.

Piana S., K. Lindorff-Larsen, D.E. Shaw, How robust are protein folding simulations with respect to force field parameterization?, *Biophys J*, 100 (2011) L47–9.

Pierce B., W. Tong, Z. Weng, M-ZDOCK: a grid-based approach for Cn symmetric multimer docking, *Bioinformatics*, 21 (2005) 1472–8.

Salmon L., M.R. Jensen, P. Bernadó, M. Blackledge, Measurement and analysis of NMR residual dipolar couplings for the study of intrinsically disordered proteins, *Methods Mol Biol*, 895 (2012) 115–25.

Santambrogio C., A.M. Frana, A. Natalello, E. Papaleo, M.E. Regonesi, S.M. Doglia, P. Tortora, G. Invernizzi, R. Grandori, The role of the central flexible region on the aggregation and conformational properties of human ataxin-3, *FEBS J*, 279 (2012) 451–63.

Scherzinger E., R. Lurz, M. Turmaine, L. Mangiarini, B. Hollenbach, R. Hasenbank, G.P. Bates, S.W. Davies, H. Lehrach, E.E. Wanker, Huntingtin-encoded polyglutamine expansions form amyloid-like protein aggregates in vitro and in vivo, *Cell*, 90 (1997) 549–558.

Song A.-X., C.-J. Zhou, Y. Peng, X.-C. Gao, Z.-R. Zhou, Q.-S. Fu, J. Hong, D.-H. Lin, H.-Y. Hu, Structural transformation of the tandem ubiquitin-interacting motifs in ataxin-3 and their cooperative interactions with ubiquitin chains, *PLoS One*, 5 (2010) e13202.

Tiberti M., G. Invernizzi, M. Lambrughi, Y. Inbar, G. Schreiber and E. Papaleo (2014). PyInteraph: a framework for the analysis of interaction networks in structural ensembles of proteins. *J Chem Inf Model* 54(5): 1537-1551.

Thakur A.K., M. Jayaraman, R. Mishra, M. Thakur, V.M. Chellgren, I.-J.L. Byeon, D.H. Anjum, R. Kodali, T.P. Creamer, J.F. Conway, A.M. Gronenborn, R. Wetzel, Polyglutamine disruption of the huntingtin exon 1 N terminus triggers a complex aggregation mechanism, *Nat Struct Mol Biol*, 16 (2009) 380–389.

Tompa P., The interplay between structure and function in intrinsically unstructured proteins, *FEBS Lett*, 579 (2005) 3346–54.

Uversky V.N., J.R. Gillespie, A.L. Fink, Why are “natively unfolded” proteins unstructured under physiologic conditions?, *Proteins*, 41 (2000) 415–27.

Uversky V.N., C.J. Oldfield, A.K. Dunker, Intrinsically disordered proteins in human diseases: introducing the D2 concept, *Annu Rev Biophys*, 37 (2008) 215–46.

Uversky V.N., Intrinsic disorder in proteins associated with neurodegenerative diseases, *Front Biosci*, 14 (2009) 5188–238.

Uversky V.N., Size-exclusion chromatography in structural analysis of intrinsically disordered proteins, *Methods Mol Biol*, 896 (2012) 179–94.

van den Heuvel R.H.H., A.J.R. Heck, Native protein mass spectrometry: from intact oligomers to functional machineries, *Curr Opin Chem Biol*, 8 (2004) 519–26.

Vishveshwara S., A. Ghosh, P. Hansia, Intra and inter-molecular communications through protein structure network, *Curr Protein Pept Sci*, 10 (2009) 146–60.

Wang Q., P. Young, K.J. Walters, Structure of S5a bound to monoubiquitin provides a model for polyubiquitin recognition, *J Mol Biol*, 348 (2005) 727–39.

Ward J.J., J.S. Sodhi, L.J. McGuffin, B.F. Buxton, D.T. Jones, Prediction and functional analysis of native disorder in proteins from the three kingdoms of life, *J Mol Biol*, 337 (2004) 635–45.

Warrick J.M., L.M. Morabito, J. Bilen, B. Gordesky-Gold, L.Z. Faust, H.L. Paulson, N.M. Bonini, Ataxin-3 suppresses polyglutamine neurodegeneration in Drosophila by a ubiquitin-associated mechanism, *Mol Cell*, 18 (2005) 37–48.

Woods L.A., G.W. Platt, A.L. Hellewell, E.W. Hewitt, S.W. Homans, A.E. Ashcroft, S.E. Radford, Ligand binding to distinct states diverts aggregation of an amyloid-forming protein, *Nat Chem Biol*, 7 (2011) 730–9.

Zhang W., D. Ganguly, J. Chen, Residual structures, conformational fluctuations, and electrostatic interactions in the synergistic folding of two intrinsically disordered proteins, *PLoS Comput Biol*, 8 (2012) e1002353.

Zhuravlev P.I., C.K. Materese, G.A. Papoian, Deconstructing the native state: energy landscapes, function, and dynamics of globular proteins, *J Phys Chem B*, 113 (2009) 8800–8812.

Zoghbi H.Y., H.T. Orr, Glutamine repeats and neurodegeneration, *Annu Rev Neurosci*, 23 (2000) 217–247.

2.3.7 Supplementary Materials

Full Supplementary Materials can be found at: **The conformational ensemble of the disordered and aggregation-protective 182-291 region of ataxin-3.** Invernizzi, G., Lambrughi, M., Regonesi, M.E., Tortora, P., Papaleo, E. (2013). *Biochim. Biophys. Acta*, 1830 (11):5236-47 http://www.sciencedirect.com/science/article/pii/S0304416513003103

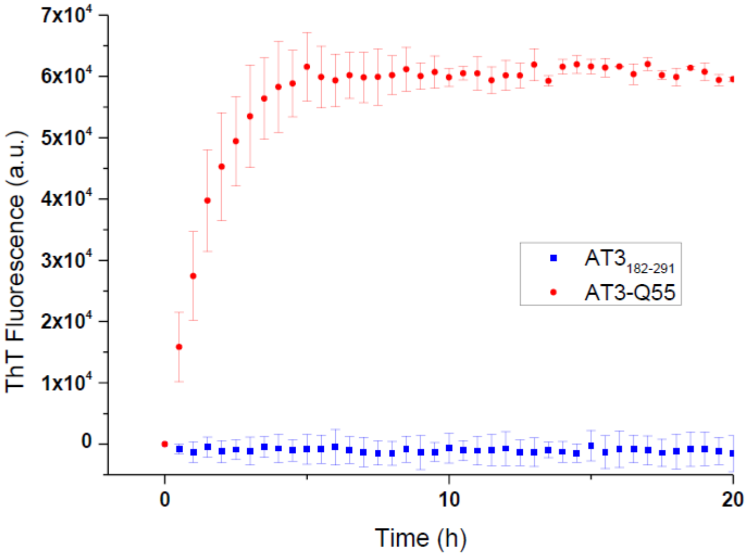


Figure S1. AT3$_{182-291}$ ThT assay. Tht fluorescence assay of 12.5 µM AT3-Q55 (red) and AT3$_{182-291}$ (blue). Error bars are relative to standard deviation calculated on at least three different experiments.

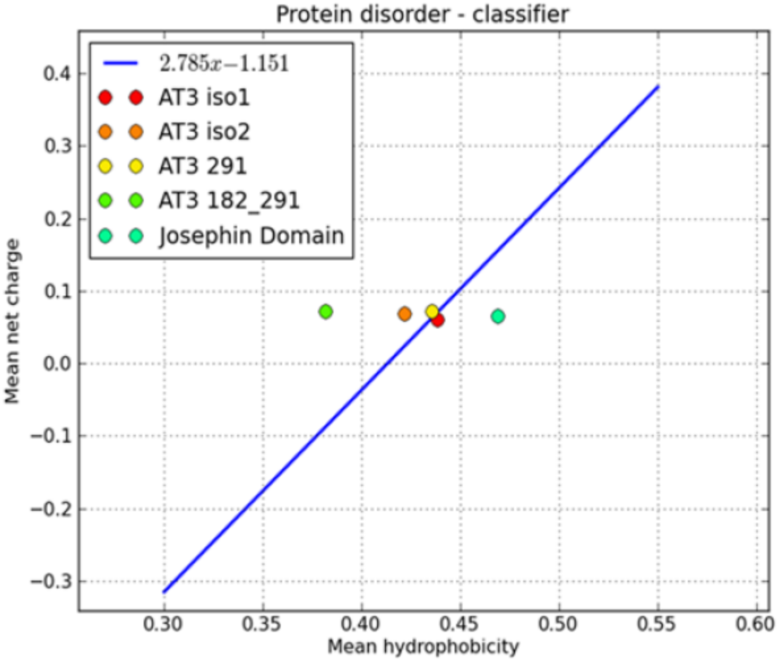


Figure S2. Two-dimensional diagram representing the mean hydrophobicity values (<H>) as a function of the mean net charge of AT3 and its domains. These two parameters are used as order indicator, as proposed by Uversky and co-workers (Uversky et al., 2000), to discriminate among folded and unfolded proteins, than in the plot, are represented as two regions of space, separated by a blue line. The green dot refers to the AT3$_{182-291}$ domain. Also values for the JD (residues 1-182, cyan dot), AT3 isoforms 1 (iso1, red dot) and 2 (iso2, orange dot) and AT3-291Δ (yellow dot) are shown

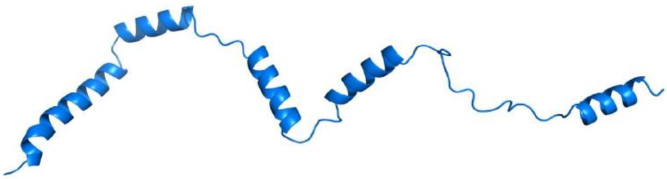


Figure S3. AT3$_{182-291}$ starting model. The starting model for MD simulations was selected among the models generated by ab-initio modeling with I-Tasser as a model that lacks sidechain-sidechain long-range intramolecular interactions.

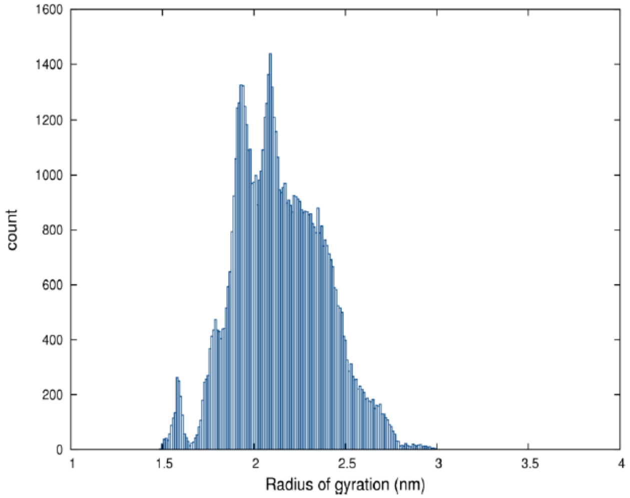


Figure S5. Distribution of the Rg values was calculated over the macro-trajectory. Two major populations are present with Rgs of 1.93 and 2.11 nm, respectively.

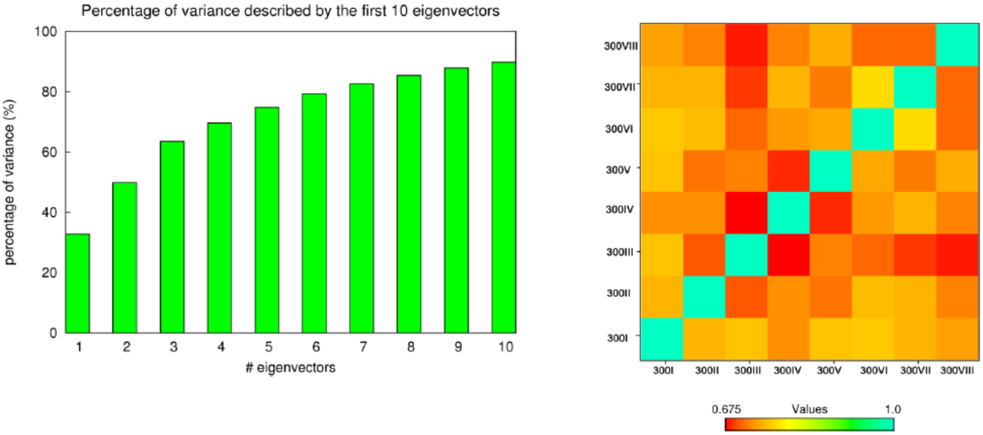


Figure S7. Evaluation of the described MD ensemble. A. Percentage of the cumulative variance described by the first ten eigenvectors. B. Rmsip matrix calculated for each pairwise comparisons of the single replicate, using the conformational space described by the first 10 principal components. The Rmsip values are highlighted by a color gradient, from cyan (1) to red (0.675).

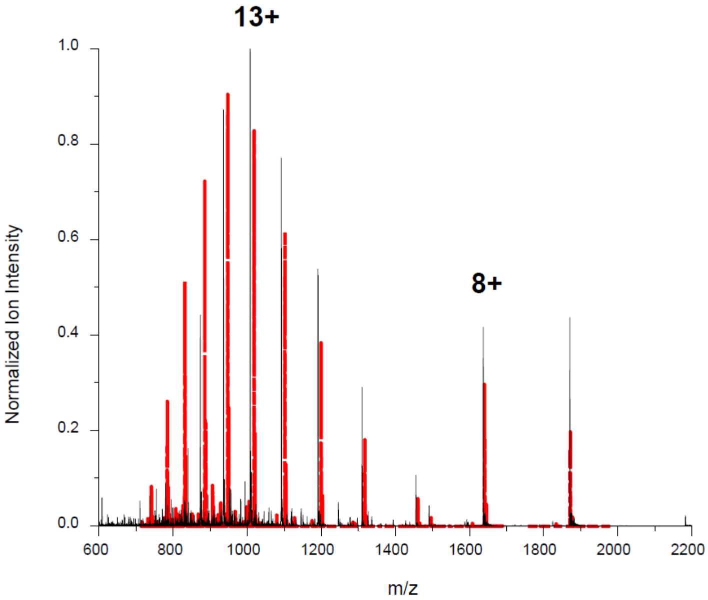


Figure S11. ESI-MS spectra of AT3182-291 at high temperature. ESI-MS spectra of 5 µM AT3$_{182-291}$ recorded at pH 6.2 with heater temperature set at 200 °C (black) and 0 °C (red).

2.4 Ensemble description of intrinsically disordered proteins by Nuclear Magnetic Resonance spectroscopy: the Ataxin3 182-291 domain

2.4.1 Introduction

The high-resolution characterization of disordered protein states is challenging due to their intrinsic heterogeneity (Dobson et al., 2003). As shown in previous chapters, MD simulations can effectively describe complex system as IDPs and characterize their dynamic properties but approximations are still present. To obtain a more consistent and solid description of the structural ensemble populated by AT3$_{182\text{-}291}$experimental data collected at atomic resolution are needed to overcome the problems releated to MD force fields. We thus exploited Nuclear Magnetic Resonance (NMR) spectroscopy (Shortle, 1996, Smith et al., 1996, Barbar, 1999, Dyson and Wright, 2004, Iešmantavičius, 2014) to achieve a deeper structural investigation and additional detailed structural information on AT3$_{182\text{-}291}$ The project has been carried out in collaboration with the Structural Biology and NMR Laboratory (SBinLab) at the Department of Biology of the University of Copenhagen (Denmark) under the supervision of Dr. Gaetano Invernizzi and Dr. Elena Papaleo in the groups of Prof. K. Teilum and Prof. K. Lindorff-Larsen in which I worked as visiting Ph.D. student from August 2013 to August 2014. Indeed, NMR spectroscopy is one of the best technique to achieve both information on structure and dynamics of proteins in solution. Indeed, NMR spectroscopy has the potential to effectively provide atomistic information on highly dynamic systems at timescales varying over many orders of magnitude (Dyson and Wright 2004). It provides both short range and long-range structural information that can be exploited to accurately describe the behavior and conformational ensemble of IDPs (Teilum et al. 2009, Teilum et al. 2011).

NMR studies on IDPs are difficult since they lack resonance dispersion in the proton dimension, resulting in overlapping signals that render difficult the resonance assignment process. Recent advances in NMR methodologies and the use of high-field NMR spectrometers permit to partially overcome this problem (Choy and Forman-Kay 2001, Dyson and Wright 2004). The achievement of anearly complete resonance assignment of protein atoms is mandatory to collect other NMR measurements to characterize a protein conformational ensemble. NMR experiments can provide data as chemical shifts, paramagnetic relaxation enhancement (PRE) from site-directed spin-labeling (Gillespie and Shortle 1997, Teilum et al. 2002, Lietzow et al. 2002, Lindorff-Larsen et al. 2004) and residual dipolar couplings (Shortle and Ackerman, 2001) that have been applied to disordered states and are effective in obtaining information about IDPs. Experimentally collected chemical shifts can be compared to average values of chemical shifts associated to random coils. The difference between the experimentally measured chemical shifts of residues and their random coil references (Merutka et al.

1995) results in secondary chemical shifts that permit to identify the regions of the protein that are characterized by secondary structures. Structural information about IDPs can be also obtained measuring residual dipolar couplings in partially aligned media by incorporation of the sample in polyacrylamide or polyethylene glycol media (Tycko et al., 2000, Shortle and Ackerman 2001). This measure permits to identify regions in the protein that assumes specific relative orientations, giving long-range information on transient tertiary structures and secondary structure elements. Long-range distance information of IDPs to complement or overcome problems related to NOEs measurements for IDPs can be obtained from paramagnetic relaxation enhancement (PRE) via covalent attachement to the protein sequence of a paramagnetic nitroxide spin label (Teilum et al. 2002, Dedmon et al. 2005). In this technique a nitroxide spin-label is attached to a specific region of the protein, usually on the thiol group of a cysteine residue on a site that do not disrupt or influence the protein structure. Nitroxide spin labels exist in two states: oxidized (paramagnetic) state with an unparied electron and reduced (diamagnetic) state. The presence of the probe in paramagnetic state enhances transversal relaxation rate of amide protons and causes the broadening of NMR nuclear signals of nearby protons, while no effects are present when the label is in diamagnetic state. PRE method is based on recording a ^{1}H-^{15}N Heteronuclear Single Quantum Coherence (HSQC) spectrum for the paramagnetic sample and a second spectrum with the spin label reduced to the diamagnetic state. The relaxation enhancement is inversely proportional to the sixth root of the distance between the nuclear spin and unpaired electron. Differences in intensities of crosspeaks in the spin labeled and reduced spectrum permit to assess the broadening effect and give an estimate of the distance of the spin label site from any given amide allowing the computation of distance information. As the effect of PRE by an unpaired electron is observable on protons at distances till 20 Å, long-range structural information can be obtained in the dynamic ensemble of highly flexible IDPs. Indeed, in addition to the local effects of the oxidized spin-label, which reduce strongly the cross-peak intensity of adjacent residues, long-range contacts, even if they are transient, between distant regions in the protein can be observed, indicating the presence of non random coil structure. We here measured some of the aforementioned NMR parameters to characterized the structural ensemble of the disordered region of human ataxin 3 AT3$_{182\text{-}291.}$.

2.4.2 Materials and Methods

Systems of expression and bacterial strains

We used BL21(DE3)-CodonPlus strain (Stratagene, La Jolla, CA, USA) that permits an efficient and high level production of heterologous proteins in *E. coli*. In fact high-level expression of eukaryotic proteins is frequently limited by the rarity of certain type of tRNAs that are abundant in humans, stalling the translation. BL21-CodonPlus strains are engineered to contain extra copies of genes *Argu*, *Iley* and *Leuw* that encode the tRNAs that recognize codons AUA, AGG, AGA, CUA, CCC and GGA rarely used in *E.coli* that most frequently limit translation of heterologous proteins. This strain is used for the expression of genes inserted into a recombinant plasmid containing the promoter for the T7 RNA polymerase. It is a lysogenic strain for the DE3 phage, derived from bacteriophage λ. The genome of this phage carries a fragment containing the *lacI* gene, the lac UV5 promoter and the gene for T7 RNA polymerase. In the lysogenic strain the only promoter capable of controlling the transcription of the T7 RNA polymerase gene is *lac UV5* by induction with lactose or isopropyl-β-D-1-thiogalactopyranoside (IPTG). In the presence of the inductor, therefore, there is the transcription of the gene of T7 RNA polymerase and the expression of this enzyme, which induces the transcription of plasmid DNA under the control of the T7 phage promoter. This system has great specificity and high selectivity since the T7 RNA polymerase binds very strongly to its promoter and can transcribe the RNA five times faster than the RNA polymerase of *E. coli*.

Plasmid cloning and expression

We used a pGEX-6p-1 plasmid as cloning and expression vector for the fragments 182-291 of human AT3 (GE Healthcare Bio -Sciences AB, Uppsala, Sweden). The pGEX vectors allow the cloning of an exogenous gene or a fragment of an aminoacid sequence in frame with the gene encoding the glutathione S-transferase (GST) used as tag for the subsequent purification procedure. In vectors pGEX-6P-1 the target genes are cloned under the control of the promoter for T7 RNA polymerase of bacteriophage λ to determine the inducible and high-level expression of the fusion protein with GST-tag in *E. coli*. The T7 RNA polymerase is highly specific for its promoter and is not expressed by the host cells, thus limiting also the basal expression of genes carried by the pGEX-6P-1. Furthermore, immediately upstream of the multiple cloning site is located a recognition sequence for the Prescission protease, a site specific protease, specific engineered with a GST tag, that allows to detach with high efficiency the GST tag from purified proteins. These plasmid vectors contain also the coding sequence of *lacI*, a bacterial origin of replication (ori) and genes for the resistance to antibiotics, such as ampicillin (Amp), used as a marker for the selection of clones of transformed E. coli.

Growth media

LB media. Medium used for the growth of *E. coli* expression strains and production of unlabeled proteins: Peptone 10g/L, Yeast Extract 5g/L, NaCl 10g/L filled to volume with MilliQ water and added with proper antibiotic, Amp 100 μg/mL

Auto-inducing growth minimal medium: medium used for the growth of *E. coli* expression strains and production of labeled proteins:

25mM Na_2HPO_4; 25mM KH_2PO_4; 5mM Na_2SO_4; 0.5% Glycerol; 0.05% Glucose; 0.2% Lactose; 100μM $FeSO_4$; 2mM $MgSO_4$; 1mM Trace Metals, 100 μM $CaCl_2$, filled to volume with MilliQ water and added with proper antibiotic, Amp 100 μg/mL

For ^{15}N labeled proteins added $^{15}NH_4Cl$ 1g/L or $(^{15}NH_4)_4SO_4$ 1g/L as sole nitrogen source.

For ^{15}N ^{13}C labeled proteins added $^{15}NH_4Cl$ 1g/L or $(^{15}NH_4)_4SO_4$ 1g/L and substituted the carbon source with a solution of 0.4% ^{13}C-glycerol and then induced with Isopropyl ß-D-1-thiogalactopyranoside (IPTG). With this minimal media *E. coli* use glucose over others carbon sources. Glucose is a strong inhibitor of lactose uptake, so till it is present in the growth media the lactose is not transported into the bacterial cells. Upon depletion of glucose, lactose enters the bacterial cells and induces the production of heterologous recombinant protein using glycerol as preferential carbon source.

Protein expression and purification

LB-medium agar plates were prepared from glycerol stocks of BL21(DE3)-CodonPlus expression strain transformed with the plasmid pGEX-6P-1 containing the gene for $AT3_{182-291}$ and $AT3_{182-291}$ mutants, M186C, T207C, S219C, M242C, S261, S278C, that were constructed by site-directed mutagenesis. A single colony was pre-inoculated in 10 ml of LB medium with addition of Amp. The growth took place overnight, with constant fast shaking (250 rpm), at 37 ° C. The next day the pre-inocule culture, with 1:500 or 1:1000 dilution, was inoculated into 1 L of auto-inducing minimal medium with Amp 100 μg / mL and incubated at 37 ° C with vigorous shaking (250 rpm) till OD600 of 0.8 and then at 30 ° C, always with constant shaking, until OD600 of 3. For the production of unlabeled proteins in LB medium at OD600 of 0.8, the expression was induced by adding 0.5 mM IPTG. Afterwards the cells were harvested by centrifugation at 7000 xg for 15 min at 4 ° C and the cellular pellets were frozen overnight.

Preparation of crude extract

The next day the cells were thawed and resuspended in 50mL of PBS-lysis buffer (50 mM KH_2PO_4, 50 mM Na_2HPO_4, 300 mM NaCl, pH 7.4) to which was added DNAse 10 μg/mL (Sigma-Aldrich, St. Louis, MO, USA) and PMSF, 1 mM. The sample was left for 30min in agitation at 4 °C till complete resuspension. The sample was sonicated for three cycles of 30s, maintaining the sample in ice. At this point sample was centrifuged at 29220 xg at 4 ° C for 45 min and supernatant collected, discharging the cellular pellet.

Affinity chromatography

The crude extract obtained was filtered and subsequently incubated with 800 μl of Glutathione Sepharose 4 Fast Flow resin (GE Healthcare Bio-Sciences, Uppsala, Sweden) for 2h at 4 °C with shaking. The resin was previously packed in a column and washed with 10 ml of MilliQ H2O and pre-equilibrated with 10 ml of PBS buffer. After collecting the excluded, the column was washed with 10 volumes of washing buffer (50 mM PBS pH 7.4, 1 mM PMSF) followed by 10 volumes of Cleavage Buffer for the PreScission Protease (50 mM PBS, pH 7.4, 150 mM NaCl, 1 mM EDTA, 1 mM DTT). The resin was then collected and incubated in Cleavage Buffer at 4 °C overnight with shaking and 60 units of PreScission Protease (HRV 3C Protease Sino Biological inc., Beijing, P.R.China) per ml of resin were used to *in-site* cleave the GST-tag. The day after the sample was reloaded on the column and eluted with PBS pH 7.4 collecting the eluted fractions. The fractions were analysed by the SDS-PAGE. The eluted samples were further purified by size-exclusion chromatography on a Superdex 75 10/300GL column (GE Healthcare LifeSciences, Little Chalfont, England).

Size-exclusion chromatography (SEC)

Size-exclusion chromatography was performed on an FPLC AKTA purifier liquid-chromatography system (Amersham Biosciences, Uppsala, Sweden) using a Superdex 75 10/300GL column (GE Healthcare LifeSciences, Little Chalfont, England). Chromatography was carried out in PBS buffer, pH 7.4, 150 mM NaCl at a flow rate of 0.5 mL/min, and monitored by absorbance at 280 nm. Thus chromatograms expressing the elution volume compared to the absorbance measured at UV 280 were obtained. All buffers were filtered and degassed. Eluted fractions of 0.5 mL were collected. The protein content was assayed by NanoDrop 1000 spectrometer and fractions monitored by SDS-PAGE electrophoresis.

NMR spectroscopy

NMR samples were prepared by dissolving purified protein in 90% PBS buffer, pH 7.4, 150 mM NaCl and 10% D_2O with4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) added as internal calibration standard. Protein concentrations were from 0.5 to 1 mM in a volume of 400 μl. Assignment of backbone chemical shifts was performed on a 0.5 mM ^{13}C,^{15}N $AT3_{182\text{-}291}$ sample and ^{1}H,^{15}N-HSQC spectrum and the following triple resonance spectra were recorded, HNCA, HN(CO)CA, HNCO, HN(CA)CO, CBCA(CO)NH, CBCANH, CC(CO)NH and H(CCO)NH (all pulse programs from Varian ProteinPack) at 4 ° C and 25 ° C on a Varian Unity Inova 750 and 800 Mhz instrument (Structural Biology and NMR Laboratory, Department of Biology, University of Copenhagen, Denmark). NMR data were processed by nmrPipe and analysed using CCPNMR.

For the measure of paramagnetic relaxation enhancement (PRE) the six site-directed cysteine mutants (M186C, T207C, S219C, M242C, S261, S278C) of $AT3_{182-291}$ were expressed and purified as stated above. The cysteine residues of the purified $AT3_{182-291}$ variants were modified with (1-oxyl-2,2,5,5-tetramethyl-3-pyrroline-3-methyl) methanesulfonate (MTSL, Toronto Research Chemicals Inc., Canada). After affinity chromatography the eluted proteins were dissolved in PBS buffer, pH 7.4, 150 mM NaCl, and a 10 M excess of MTSL from a 200 mM stock in ethanol was added. The reaction was left overnight in the dark, under gentle shaking and at 4 ° C then the $AT3_{182-291}$ variants were purified by FPLC. The homogeneity of the purified spin-labelled protein was verified by MALDI-TOF MS.

NMR samples were prepared by dissolving spin-labelled protein in 90% PBS buffer, pH 7.4, 150 mM NaCl, 10% D_2O and DSS was added as calibration standard. Protein concentrations were 0.3 mM in a volume of 400 µl. $^1H,^{15}N$-HSQC NMR spectra were recorded on a Varian Unity Inova 750 MHz spectrometer at 4 ° C and 25 ° C (Structural Biology and NMR Laboratory, Department of Biology, University of Copenhagen, Denmark). Spectra were recorded on each sample both before (oxidized or paramagnetic state) and after reduction (reduced or diamagnetic state) of the spin label with a 2-fold excess of ascorbate added from a 1 M stock and adjusted to the pH of 7.4.

2.4.3 Results

Even if we have provided a first structural characterization of $AT3_{182-291}$ fragment, which is described in Chapter 2.3, we are conscious of the limits of canonical MD simulations of IDPs and in the low resolution biophysical techniques that have accompanied the previous computational study. We thus exploit NMR to achieve a deeper and more accurate description of the structural properties of $AT3_{182-291}$ (Figure 1). Firstly we identify the best conditions and protocols to produce, express and purify at high yield the isotopically labeled $AT3_{182-291}$ in *E. coli*, obtaining the ^{15}N and ^{13}C-^{15}N labeled protein at a NMR grade i.g. at a concentration of 0.5 mM. $AT3_{182-291}$ was expressed in *E.Coli* BL21 strain in a PGEX-6P-1 plasmid cloned in fusion with glutathione S-transferase (GST).

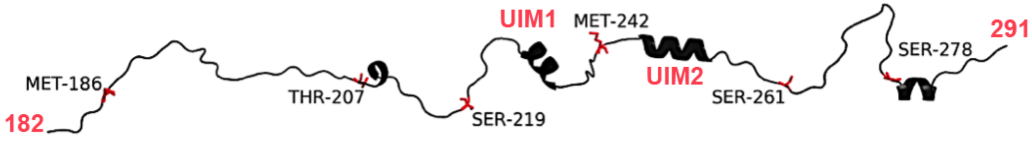


Figure 1. Cartoon representation of an extended model of the AT3$_{182\text{-}291}$. The residues mutated to cysteine and modified with spin label are shown in stick representation.

The protein was purified as described in Materials and Methods and analysed by SDS-PAGE (Figure 2). We also produced six single point AT3$_{182\text{-}291}$ mutants by site-direct mutagenesis to perform specific NMR experiments (Figure 1). The NMR experiments were performed using a Variant Unity Inova 750 and 800 MHz spectrometers. We initially collected ^{1}H-^{15}N Heteronuclear Single Quantum Coherence (HSQC) spectra at 4 ° C and 25 ° C. The ^{15}N-^{1}H HSQC spectra pointed out a low chemical shifts dispersion in the proton dimension, confirming that AT3$_{182\text{-}291}$ is mainly disordered in solution (Figure 2).

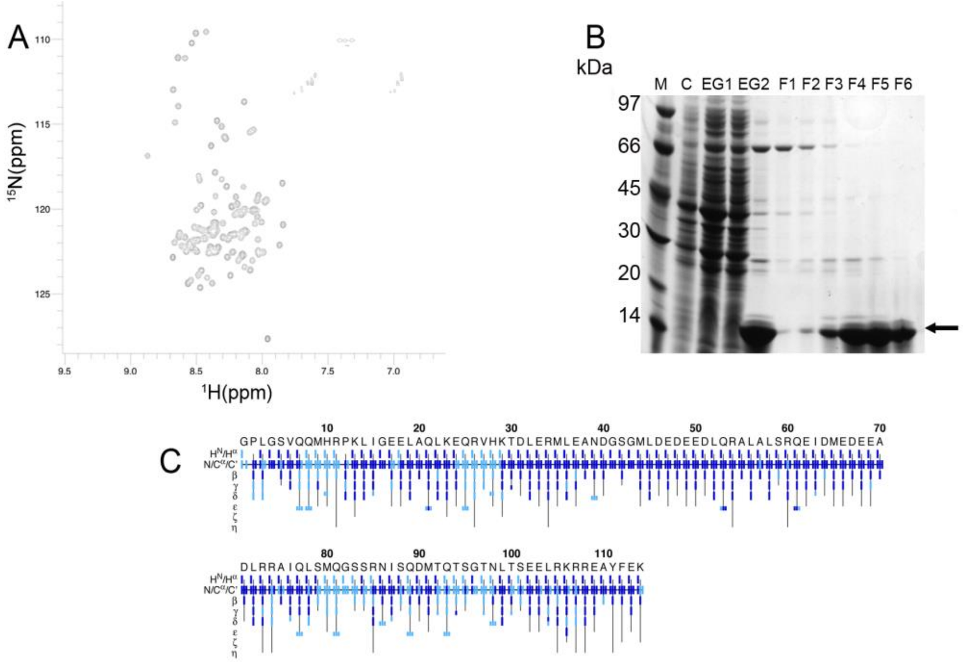


Figure 2. ^{15}N-^{1}H HSQC spectra of wild-type AT3$_{182\text{-}291}$ in PBS buffer, 150 mM NaCl, pH 7.4 at 4 ° C. A) ^{15}N-^{1}H HSQC spectra at 4 ° C. The spectra at 4 ° C and 25 ° C are very similar but variations are present in the 25 ° C spectrum and some resonance signals disappear (*data not shown*). B) SDS-PAGE of the purification of the AT3$_{182\text{-}291}$: crude extract (EG1, 2 50 ml total), eluted fraction obtained after

cleavage with Prescission Protease (1 ml total), eluted fractions (F1-F6, 1 ml total each) obtained after purification with affinity chromatography and SEC. M indicates protein molecular weight markers (LMW-SDS Marker Kit, GE). C) Combining HSQC and several triple resonance NMR spectra the ^{13}Cα, ^{13}CO, ^{15}N, and ^{1}HN backbone chemical shifts for all residues (99% of the atoms at 4 ° C) and side chain for the majority of the residues (90% of the atoms at 4 ° C) were assigned under 4 ° C and 25 ° C.

We proceeded to assign the protein by mean of several different complementary triple-resonance experiments on ^{13}C-^{15}N labeled forms, like HNCA, HN(CO)CA, HNCO, HN(CA)CO, CBCA(CO)NH, CBCANH, CC(CO)NH and H(CCO)NH spectra, that together permitted us to provide sequential linking between resonances and almost complete protein assignment (Figure 2). Combining ^{15}N-^{1}H HSQC and several triple resonance NMR spectra, ^{13}Cα, ^{13}CO, ^{15}N, ^{1}Hα and ^{1}HN backbone chemical shifts for all residues were assigned under 4 ° C and 25° C. Moreover side chain ^{13}C for the majority of the residues, except for the last carbon atoms on long side chains, were assigned under 4 ° C. Since the chemical shifts of backbone Cα and CO and, to a smaller extent, Cβ are dependent on secondary structure (Wishart et al. 1992) they can be analysed to estimate the propensity of AT3$_{182\text{-}291}$ to populate secondary structure conformations. Therefore we used δ2D a software that permits to obtain probabilistic information about the distribution of populations of secondary structure elements from the backbone chemical shifts of disordered proteins (Camilloni et al. 2012) analyzing the distinct deviations from random coil chemical shifts values. From the analysis of the assigned chemical shifts by δ2D it turned out that (Figure 3) large part of AT3$_{182\text{-}291}$, especially the N-terminal region, has random coil behavior. Nevertheless, we identified at least four regions of AT3$_{182\text{-}291}$ with high propensity to populate helical conformations: region I Asp208-Met212, region II Leu227-Leu235, region III Glu245-Leu255, region IV Ser278-Arg285. Two of these tracts, region II and region III overlap with the two UIMs, located Glu224-Ile240 (UIM1) and Asp244-Ser260 (UIM2). The NMR-derived structure of the complex between a construct bearing the two UIMs motifs of AT3 (residues 222-263 PDB entry 2KLZ, Song et al. 2010) show that the two UIMs pair together and form two α-helices. We here show that the UIM motifs of AT3$_{182\text{-}291}$ are in helical conformation evenin absence of Ubiqutin (Ub). These results are also in agreement with the secondary structure content observed in our published MD ensemble (Invernizzi et al. 2013).

To investigate if some long-range interactions, also transient, are present in the AT3$_{182\text{-}291}$ fragment and to obtain information on possible tertiary contact, we used site-directed spin-labelling and performed paramagnetic relaxation enhancement (PRE) experiments.

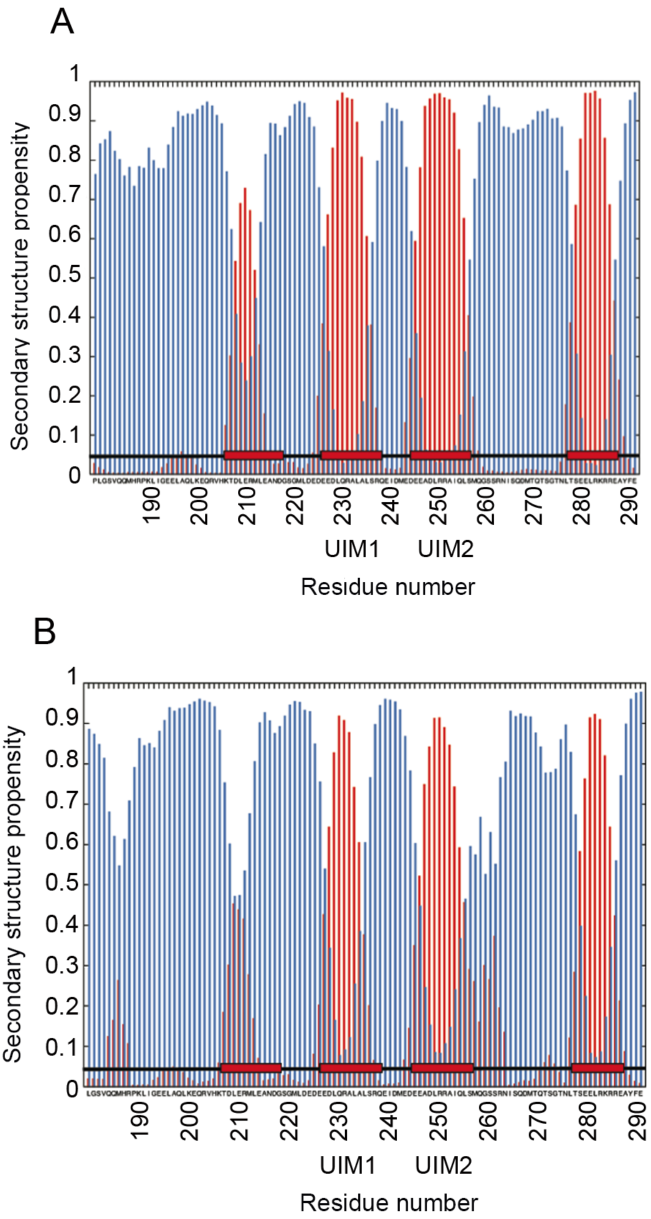


Figure 3. Chemical shift analysis of ^{13}Cα, ^{13}Cβ and ^{13}CO nuclei in AT3$_{182\text{-}291}$ performed by δ2D. Chemical shifts measured at 4 °C (A) and 25 °C (B) were compared to chemical shifts of the amino acid residues in unstructured peptides. The locations of the four helices predicted by the previously performed MD simulations are indicated with red boxes. The position of the two UIMs is indicated.

We used the thiol-specific spin-label probe MTSL *S*-(1-oxyl-2,2,5,5 tetramethyl-2,5-dihydro-1H-pyrrol-3-yl)methyl methanesulfonothioate (MTSL) that when incubated with proteins forms disulfide bonds with the thiol group in the side chain of cysteine permitting site-direct labelling. Since in the $AT3_{182\text{-}291}$ no cysteines are present we produced by site-direct mutagenesis six variants each with a different cysteine mutation along the fragment: M186C, T207C, S219C, M242C, S261C, S278C (Figure 1). The mutant variants of $AT3_{182\text{-}291}$ have been expressed in *E. coli* as isotopically labeled ^{15}N protein and purified at high yield at a NMR grade in the same way as the wild-type $AT3_{182\text{-}291}$, as described in materials and methods and analysed by SDS-PAGE (Figure 4). After purification and affinity chromatography in the $AT3_{182\text{-}291}$ mutant variants the thiol group of newly introduced cysteines were modified with paramagnetic spin-label MTSL. In order to ensure efficient labelling, the $AT3_{182\text{-}291}$ variants were labelled specifically with a tenfold excess of MTSL at pH 7.4 overnight in the dark at 4 ° C and almost complete labelling of the target protein was achieved. After purification by size exclusion chromatography a homogeneous labeled sample was obtained, as judged by matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF MS). Moreover SEC point out a monomeric state for the $AT3_{182\text{-}291}$–MTSL variants showing that the observed effects were due to intramolecular long-range contacts, rather than from dimerization or degradation and aggregation. The NMR experiments were performed using a Variant Unity Inova 750 spectrometer.

To obtain a general view of the structural properties of $AT3_{182\text{-}291}$ mutants in solution, we collected ^{15}N-^{1}H HSQC at different temperature (4 ° C and 25 ° C) of the $AT3_{182\text{-}291}$ mutants after labeling with MTSL and reduction to its diamagnetic states, switching off the PRE effect. We compared the ^{15}N-^{1}H HSQC spectrum of wild-type (wt) $AT3_{182\text{-}291}$ with the spectra of its mutant forms to check that the mutations to cysteine residues and the introduction of the spin-label to different sites in the protein had no effect on the structural ensemble of the protein . The comparison shows that only little alterations can be observed in comparison with the wt $AT3_{182\text{-}291}$ for all the mutants and changes are mostly present only in the residues closer to the labeling site, except partially for the S261C mutant that is discussed in details below, supporting the feasibility of the approach. The assignment of the ^{15}N-^{1}H HSQC spectra of the six MTSL-labelled mutants was achieved by comparison with the ^{15}N-^{1}H HSQC spectrum wt $AT3_{182\text{-}291}$ from which the assignment was transferred. The PRE effect is based on the presence of the spin-label probe MTSL attached to the single cysteine of each $AT3_{182\text{-}291}$ variant, that is located in a known region of the protein.

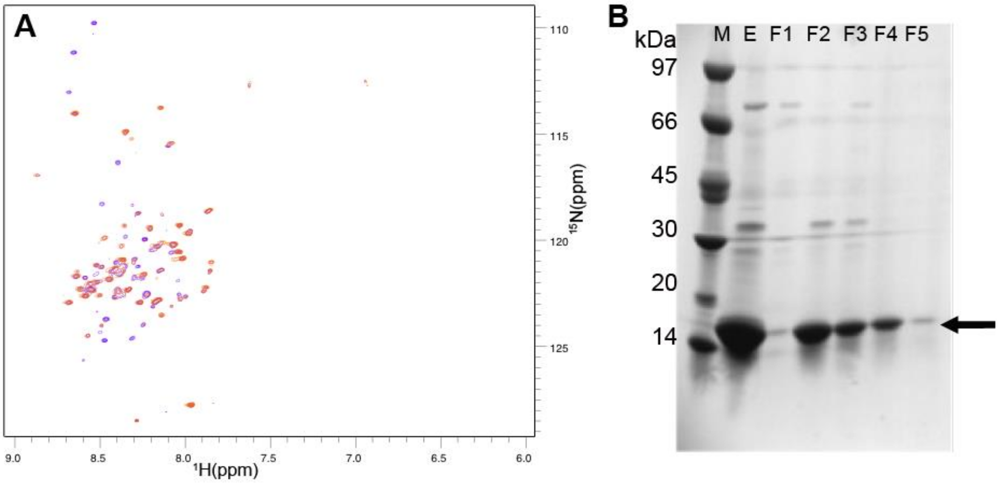


Figure 4. ^{15}N-^{1}H HSQC spectra of spin-labelled AT3$_{182\text{-}291}$ M186C-MTSL in PBS buffer, 150 mM NaCl, pH 7.4 pH 7.4 at 4 ° C. A) The spin label was reduced with ascorbate prior to acquisition and the ^{15}N-^{1}H HSQC spectra (violet) compared to the ^{15}N-^{1}H HSQC collected with the spin label in paramagnetic state (orange). The spectra are very similar without large variations of the signals as in the native-state. B) SDS-PAGE of the purification of the spin-labelled AT3$_{182\text{-}291}$ M186C-MTSL: eluted fraction obtained after cleavage with Prescission Protease (E 1 ml total), eluted fractions (F1-F5, 1 ml total each) obtained after purification with affinity chromatography and SEC. M indicates protein molecular weight markers (LMW-SDS Marker Kit, GE).

The MTSL in oxidated state has a nitroxide radical (paramagnetic state) with an unpaired electron that interact with nearby protons causing broadening of their NMR signals by an increase in transverse relaxation rate. MTSL can be reduced by reducing agent as ascorbic acid and in this state (diamagnetic state) its effect on NMR signals is switched off. The effect of the spin label and the strength of the interactions between the paramagnetic spin label MTSL and nearby protons is measured by the decrease in the intensities of ^{15}N-^{1}H HSQC cross-peaks in the paramagnetic sample (Iox) relative to the intensities in the diamagnetic sample (Ired) and is quantified using the intensity ratio Iox/Ired. The PRE effect is dependent on the distance between nitroxide radical electron of MTSL and protons of protein and is observable on resonances intensities of nuclei at distances up to 20 Å, making possible to collect long-range structural information in the dynamic ensemble of highly flexible disordered proteins (Battiste and Wagner 2000). We measured the peak intensities in ^{15}N-^{1}H HSQC NMR spectra with the probe in paramagnetic and its diamagnetic form, after reduction with 2M excess of ascorbic acid, obtaining distance information between the spin-label and the amide protons in AT3$_{182\text{-}291}$.

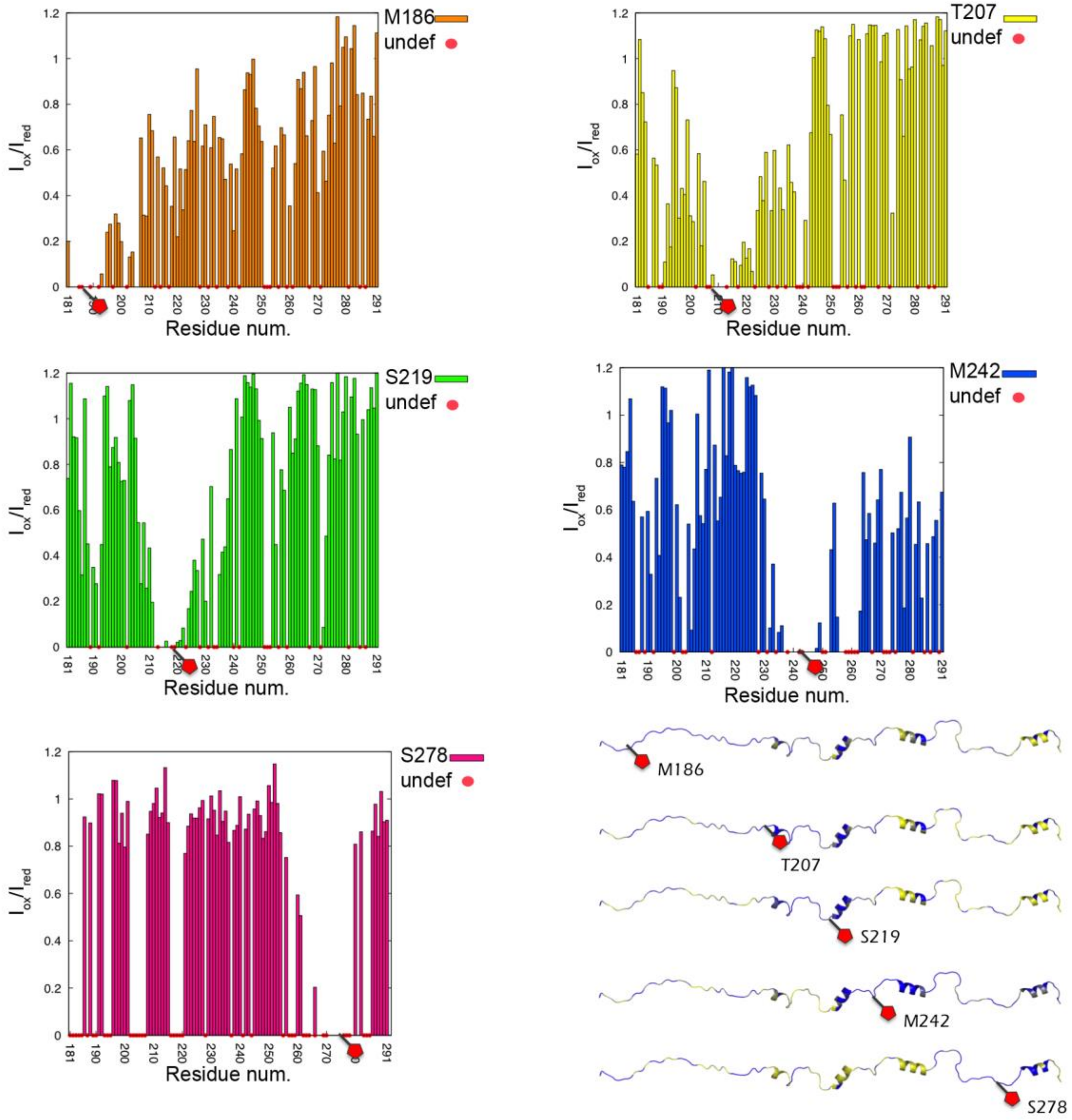


Figure 5. PRE of backbone amide protons in spin-labeled AT3$_{182-291}$ under native conditions in PBS buffer, 150 mM NaCl, pH 7.4 at 4 ° C. Six Cys-substituted and spin-labelled variants of AT3$_{182-291}$ M186C-MTSL, T207C-MTSL, S219C-MTSL, M242C-MTSL, S261-MTSL and S278C-MTSL were analysed. The HSQC spectra of the protein sample before (oxidated, paramagnetic) and after (reduced, diamagnetic) reduction of the spin label with ascorbate were recorded and the intensity ratio of the HN,N cross-peaks between paramagnetic sample (Iox) relative to diamagnetic sample (Ired)

were determined. The lollipop symbols indicate the position of the modified amino acid residue and the spin-label site. White spaces and red spots in the diagrams are due to the impossibility to univocally measure the peak height as a consequence of overlapping or non-assigned residues. The intensities measured are showed on a cartoon model of AT3$_{182-291}$ in which region non affected (yellow) or affected (blue) by the presence of the probe are shown.

We used the intensity ratio between spectra with oxidated and reduced MTSL as an indicator of interactions: a ratio close to zero indicates a measurable paramagnetic effect and strong electron-proton interaction while a ratio close to 1 indicates no PRE effects and thus an average distance larger than ~20 Å.

The analysis shows that in all the mutants of AT3$_{182-291}$, resonance intensities from nuclei close to the spin label are affected in a distance-dependent manner. Effects can be also observed far beyond the nuclei in proximity of the spin-label site. These long-range effects suggest that the labeled regions make contact at less than 20 Å with other regions of the protein. Moreover as observed for other disordered proteins (Teilum et al., 2002), around the sites of the spin labels the increased relaxation effect extends far along the sequence, confirming that AT3$_{182-291}$ is highly flexible and disordered. Especially for the AT3$_{182-291}$ M186C-MTSL long range PRE effects can be observed along a large part of the sequence (till residue 270) suggesting that the N-terminal region (residue 180-200) is high flexible and behaves like a random coil, populating several different conformations and making transient interactions with different regions of AT3$_{182-291}$. The comparison of the PRE profiles collected for the AT3$_{182-291}$ M186C-MTSL, T207C-MTSL and S219C-MTSL pointed out that at least 3 different regions of AT3$_{182-291}$ are affected by the presence of the probe (residues 185-193, 207-220 and 250-260) suggesting that these regions can make long range contacts (Figure 5). This complementarity in the PRE effects measured for the three different mutant constructs and the similarity in their HSQC spectra support the reliability of the observed interactions, confirming that they are not induced by the mutations. Also the M242C-MTSL supports the presence of the long-range contacts identified before but the location of the probe in the middle of the two UIMs partially alters the dynamics in this region and shifts some of the peaks in the ^{15}N-^{1}H HSQC spectra, making difficult to perform a deep analysis of the result. The S278C-MTSL does not show clear effects in other region of the fragment showing that the C-terminus (considered from residue 270 to 291) is the part of the protein less affected by the paramagnetic elements. These evidences suggest that the C-terminal region of AT3$_{182-291}$ harbors a region with considerable helical propensity but makes only few contacts with the rest of the molecule probably behaving as extended region in its unbound state. The S261C-MTSL have an ^{15}N-^{1}H HSQC spectra with a high number of overlapping and shifted peaks in comparison to the one of wt AT3$_{182-291}$ thus resulting problematic to analyse. The S261 is located very close to the UIM2 and the introduction of the mutations and the spin label probably alter the dynamics of the protein. Interesting the regions involved in the long range contacts comprise part of the two UIMs suggesting that also in the absence of ubiquitin the two UIMs can pair together (Figure 5). The presence of long-range contacts suggests that AT3$_{182-291}$ is not completely random coil but can also populate an ensemble of more compact species.

2.4.4 Discussion

Proteins involved in the development of polyglutamine (polyQ) diseases are generally complex multidomain proteins that are characterized by both folded and disordered domains. Their biological functions and the molecular mechanisms responsible of their aggregation upon polyQ expansion are still matter of debate. In this context the characterization of structural properties and dynamics of polyQ protein is first but important step towards the understanding of their functional and disfunctional beahaviour. Ataxin-3 is a multidomain polyQ protein that presents the N-terminal structured Josephin domain (JD) (Ellisdon et al., 2006) and a C-terminal disordered region comprising the AT3$_{182-291}$, PolyQ and another short tract after the polyQ. The region of AT3$_{182-291}$ spanning from JD to polyQ contains two Ubiquitin Interacting Motif (UIM) but its role into the physiologically functions of AT3 is not clear. The AT3$_{182-291}$ is particularly important since it has been demonstrated its key role in the aggregation process of AT3, in agreement with structural investigation pointing out that polyQ flanking regions can play a role in modulating the fibrillogenesis and aggregation of the whole protein and, in turn, the onset and the toxicity of the disease (Duennwald et al., 2006, de Chiara et al., 2005, Robertson et al., 2008, Saunders and Bottomley, 2009. Recent works showed that AT3$_{182-291}$ greatly increases the aggregation rate of the JD (Santambrogio et al., 2012) and its presence, in the absence of the polyQ, is toxic and induces neurological symptoms in a murine model (Hubener et al., 2011). Furthermore a recent work, carried out in our group, points out a protective effect against aggregation of polyQ proteins by AT3$_{182-291}$. In this light, the structural characterization of AT3 disordered tract is critical to understand the molecular mechanism under AT3 aggregation and toxicity. We previously provided a first model of the conformational ensemble of the isolated 182-291 fragment by atomistic explicit solvent multi-replicate MD simulations integrated to low resolution biophysical experimental data (Invernizzi et al., 2013). In this framework, NMR spectroscopy is a suitable technique to achieve a deeper and more detailed structural investigation of the AT3 disordered region. Indeed, NMR spectroscopy can describe even highly dynamic and heterogeneous systems (Schneider et al., 2012) and provides parameters that are probes of both short range and long-range interactions, even transient in nature, making it one of the best available biophysical spectroscopies to study IDPs (Kjaergaard and Poulsen, 2012, Jensen et al., 2009, Salmon et al., 2010). The NMR characterization of AT3$_{182-291}$

reveals that a major part of this region is intrinsically disordered.Moreover the analysis of the chemical shifts, confirmed by residual dipolar couplings (RDCs) collected in our group, point out at least four different regions in the AT3$_{182\text{-}291}$ with considerable propensity to populate helical conformations. Two of these tracts correspond to the two UIMs showing that they have strong propensity to helical conformations when AT3$_{182\text{-}291}$ is free in solution. In the presence of ubiquitin the two UIMs pair together and form two α-helices (Song et al. 2010). Our NMR data show that the two UIMs for the majority of their structural ensemble are stable α-helices also in the unbound state..

Moreover, we performed paramagnetic relaxation enhancement (PRE) to obtain long-range information on the AT3$_{182\text{-}291.}$ The PRE effects extend to sites far from the attached spin label, suggesting the presence of transient long range interaction between different part of the protein. The comparison of the PRE profiles collected for the six AT3$_{182\text{-}291}$-MTSL mutants allows to identify that at least three different regions of AT3$_{182\text{-}291}$ can be involved in long range contacts (residues 185-193, 207-220 and 250-260) comprising also the two UIMs. Moreover, the experimental data suggest that AT3$_{182\text{-}291}$ C-terminal part (residues 270 to 291) does not form long range interaction with the rest of the protein, as indicated by the small effect of spin label when attached at Ser278 site. We can speculate that part of these long range effects can be interpreted as transient interaction between UIM1 and UIM2.

The existence of helical UIMs in the free state of the fragment support the involvement of AT3$_{182\text{-}291}$ UIM motifs in the recruitment of polyubiquitinated substrates. The UIM motifs are likely to cooperate with the folded Josephin domain (JD) to process polyUb chains. They might be useful elements to position the polyubiquitin chain close to the JD catalytic site and allow AT3 to sequentially perform deubiquitinating activity on the polyubiquitin chain from distal to proximal monomers. In conclusion, we here achieved the first structural characterization of the whole disordered region in AT3 by NMR spectroscopy, which allow to complement the previous computational study and to critically evaluate the structures sampled by the MD simulations carried out with the GROMOS96 43a1 force field. In particular, we demonstrated that transient long-range interactions, also interesting the UIM motifs, and segments with a high propensity for helical structure are present in the unbound state of AT3$_{182\text{-}291}$, opening new directions for further studies on their role for AT3 biological function. Overall, the experimental data collected at atomic resolution by NMR and other data that were available on the system point out that the AT3$_{182\text{-}291}$ in solution is more extended and has very little propensity to populate collapsed states comparable to the ones that we observed by ESI-MS experiments and from the MD simulations collected in the first work. This observation highlights an important limitation of current MD force fields for the study of IDPs, which always encounter the risk of overcompation of the ID domain under investigation and thus make even more important the integration of NMR and MD to study the IDP heterogeneous conformational landscape.

2.4.5 References

Bai J.J., S. S. Safadi, P. Mercier, K. R. Barber, G. S. Shaw, (2013), Ataxin-3 Is a Multivalent Ligand for the Parkin Ubl Domain. *Biochemistry* 52, 7369-7376.

Barbar, E. (1999). NMR characterization of partially folded and unfolded conformational ensembles of proteins. *Biopolymers* 51(3): 191-207.

Battiste, J. L. and G. Wagner (2000). Utilization of site-directed spin labeling and high-resolution heteronuclear nuclear magnetic resonance for global fold determination of large proteins with limited nuclear overhauser effect data. *Biochemistry* 39(18): 5355-5365.

Bernado, P., L. Blanchard, P. Timmins, D. Marion, R. W. Ruigrok and M. Blackledge (2005). A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proc Natl Acad Sci U S A* 102(47): 17002-17007.

Burnett, B., F. Li and R. N. Pittman (2003). The polyglutamine neurodegenerative protein ataxin-3 binds polyubiquitylated proteins and has ubiquitin protease activity. *Hum Mol Genet* 12(23): 3195-3205.

Camilloni, C., A. De Simone, W. F. Vranken and M. Vendruscolo (2012). Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. *Biochemistry* 51(11): 2224-2231.

Choy, W. Y. and J. D. Forman-Kay (2001). Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *J Mol Biol* 308(5): 1011-1032.

De Chiara, C., R. P. Menon, F. Dal Piaz, L. Calder and A. Pastore (2005). Polyglutamine is not all: the functional role of the AXH domain in the ataxin-1 protein. *J Mol Biol* 354(4): 883-893.

Dedmon, M. M., K. Lindorff-Larsen, J. Christodoulou, M. Vendruscolo and C. M. Dobson (2005). Mapping long-range interactions in alpha-synuclein using spin-label NMR and ensemble molecular dynamics simulations. *J Am Chem Soc* 127(2): 476-477.

Dill, K. A. and D. Shortle (1991). Denatured states of proteins. *Annu Rev Biochem* 60: 795-825.

Dobson, C. M. (2003). Protein folding and misfolding. *Nature* 426(6968): 884-890.

Duennwald, M. L., S. Jagadish, P. J. Muchowski and S. Lindquist (2006). Flanking sequences profoundly alter polyglutamine toxicity in yeast. *Proc Natl Acad Sci U S A* 103(29): 11045-11050.

Durcan, T. M., M. Kontogiannea, T. Thorarinsdottir, L. Fallon, A. J. Williams, A. Djarmati, T. Fantaneanu, H. L. Paulson and E. A. Fon (2011). The Machado-Joseph disease-associated mutant form of ataxin-3 regulates parkin ubiquitination and stability. *Hum Mol Genet* 20(1): 141-154.

Dyson, H. J. and P. E. Wright (2004). Unfolded proteins and protein folding studied by NMR. *Chem Rev* 104(8): 3607-3622.

Ellisdon, A. M., B. Thomas and S. P. Bottomley (2006). The two-stage pathway of ataxin-3 fibrillogenesis involves a polyglutamine-independent step. *J Biol Chem* 281(25): 16888-16896.

Fiebig K. M., H. Schwalbe, M. Buck, L. J. Smith, C. M. Dobson,(1996), Toward a Description of the Conformations of Denatured States of Proteins. Comparison of a Random Coil Model with NMR Measurements. *The Journal of Physical Chemistry* 100, 2661-2666.

Gillespie, J. R. and D. Shortle (1997). Characterization of long-range structure in the denatured state of staphylococcal nuclease. I. Paramagnetic relaxation enhancement by nitroxide spin labels. *J Mol Biol* 268(1): 158-169.

Hubener, J., F. Vauti, C. Funke, H. Wolburg, Y. Ye, T. Schmidt, K. Wolburg-Buchholz, I. Schmitt, A. Gardyan, S. Driessen, H. H. Arnold, H. P. Nguyen and O. Riess (2011). N-terminal ataxin-3 causes neurological symptoms with inclusions, endoplasmic reticulum stress and ribosomal dislocation. *Brain* 134(Pt 7): 1925-1942.

Iesmantavicius, V. et al., 2014, Helical propensity in an intrinsically disordered protein accelerates ligand binding. *Angew Chem Int Ed Engl*, 2014. 53(6): 1548-51.

Invernizzi, G., M. Lambrughi, M. E. Regonesi, P. Tortora and E. Papaleo (2013). The conformational ensemble of the disordered and aggregation-protective 182-291 region of ataxin-3. *Biochim Biophys Acta* 1830(11): 5236-5247.

Jensen, M. R., P. R. Markwick, S. Meier, C. Griesinger, M. Zweckstetter, S. Grzesiek, P. Bernado and M. Blackledge (2009). Quantitative determination of the conformational properties of partially folded and intrinsically disordered proteins using NMR dipolar couplings. *Structure* 17(9): 1169-1185.

Kjaergaard, M. and F. M. Poulsen (2012). Disordered proteins studied by chemical shifts. *Prog Nucl Magn Reson Spectrosc* 60: 42-51.

Kobayashi, T., K. Tanaka, K. Inoue and A. Kakizuka (2002). Functional ATPase activity of p97/valosin-containing protein (VCP) is required for the quality control of endoplasmic reticulum in neuronally differentiated mammalian PC12 cells. *J Biol Chem* 277(49): 47358-47365.

Kortemme, T., M. J. Kelly, L. E. Kay, J. Forman-Kay and L. Serrano (2000). Similarities between the spectrin SH3 domain denatured state and its folding transition state. *J Mol Biol* 297(5): 1217-1229.

Lietzow, M. A., M. Jamin, H. J. Dyson and P. E. Wright (2002). Mapping long-range contacts in a highly unfolded protein. *J Mol Biol* 322(4): 655-662.

Lindorff-Larsen, K., S. Kristjansdottir, K. Teilum, W. Fieber, C. M. Dobson, F. M. Poulsen and M. Vendruscolo (2004). Determination of an ensemble of structures representing the denatured state of the bovine acyl-coenzyme a binding protein. *J Am Chem Soc* 126(10): 3291-3299.

Mao, Y., F. Senic-Matuglia, P. P. Di Fiore, S. Polo, M. E. Hodsdon and P. De Camilli (2005). Deubiquitinating function of ataxin-3: insights from the solution structure of the Josephin domain. *Proc Natl Acad Sci U S A* 102(36): 12700-12705.

Merutka, G., H. J. Dyson and P. E. Wright (1995). 'Random coil' 1H chemical shifts obtained as a function of temperature and trifluoroethanol concentration for the peptide series GGXGG. *J Biomol NMR* 5(1): 14-24.

Mishra, A., P. Dikshit, S. Purkayastha, J. Sharma, N. Nukina and N. R. Jana (2008). E6-AP promotes misfolded polyglutamine proteins for proteasomal degradation and suppresses polyglutamine protein aggregation and toxicity. *J Biol Chem* 283(12): 7648-7656.

Robertson, A. L., J. Horne, A. M. Ellisdon, B. Thomas, M. J. Scanlon and S. P. Bottomley (2008). The structural impact of a polyglutamine tract is location-dependent. *Biophys J* 95(12): 5922-5930.

Salmon, L., G. Nodet, V. Ozenne, G. Yin, M. R. Jensen, M. Zweckstetter and M. Blackledge (2010). NMR characterization of long-range order in intrinsically disordered proteins. *J Am Chem Soc* 132(24): 8407-8418.

Santambrogio, C., A. M. Frana, A. Natalello, E. Papaleo, M. E. Regonesi, S. M. Doglia, P. Tortora, G. Invernizzi and R. Grandori (2012). The role of the central flexible region on the aggregation and conformational properties of human ataxin-3. *FEBS J* 279(3): 451-463.

Saunders, H. M. and S. P. Bottomley (2009). Multi-domain misfolding: understanding the aggregation pathway of polyglutamine proteins. *Protein Eng Des Sel* 22(8): 447-451.

Schneider, R., J. R. Huang, M. Yao, G. Communie, V. Ozenne, L. Mollica, L. Salmon, M. R. Jensen and M. Blackledge (2012). Towards a robust description of intrinsic protein disorder using nuclear magnetic resonance spectroscopy. *Mol Biosyst* 8(1): 58-68.

Shortle, D. and M. S. Ackerman (2001). Persistence of native-like topology in a denatured protein in 8 M urea. *Science* 293(5529): 487-489.

Shortle, D. R. (1996). Structural analysis of non-native states of proteins by NMR methods. *Curr Opin Struct Biol* 6(1): 24-30.

Smith, L. J., K. M. Fiebig, H. Schwalbe and C. M. Dobson (1996). The concept of a random coil. Residual structure in peptides and denatured proteins. *Fold Des* 1(5): R95-106.

Song, A. X., C. J. Zhou, Y. Peng, X. C. Gao, Z. R. Zhou, Q. S. Fu, J. Hong, D. H. Lin and H. Y. Hu (2010). Structural transformation of the tandem ubiquitin-interacting motifs in ataxin-3 and their cooperative interactions with ubiquitin chains. *PLoS One* 5(10): e13202.

Teilum, K., B. B. Kragelund and F. M. Poulsen (2002). Transient structure formation in unfolded acyl-coenzyme A-binding protein observed by site-directed spin labelling. *J Mol Biol* 324(2): 349-357.

Teilum, K., J. G. Olsen and B. B. Kragelund (2009). Functional aspects of protein flexibility. *Cell Mol Life Sci* 66(14): 2231-2247.

Teilum, K., J. G. Olsen and B. B. Kragelund (2011). Protein stability, flexibility and function. *Biochim Biophys Acta* 1814(8): 969-976.

Teilum, K., F. M. Poulsen and M. Akke (2006). The inverted chevron plot measured by NMR relaxation reveals a native-like unfolding intermediate in acyl-CoA binding protein. *Proc Natl Acad Sci U S A* 103(18): 6877-6882.

Teilum, K., M. H. Smith, E. Schulz, L. C. Christensen, G. Solomentsev, M. Oliveberg and M. Akke (2009). Transient structural distortion of metal-free Cu/Zn superoxide dismutase triggers aberrant oligomerization. *Proc Natl Acad Sci U S A* 106(43): 18273-18278.

Tollinger, M., N. R. Skrynnikov, F. A. Mulder, J. D. Forman-Kay and L. E. Kay (2001). Slow dynamics

Tycko R., F. J. Blanco, Y. Ishii, (2000), Alignment of Biopolymers in Strained Gels:? A New Way To Create Detectable Dipole-Dipole Couplings in High-Resolution Biomolecular NMR. Journal of the American Chemical Society 122, 9340. in folded and unfolded states of an SH3 domain. *J Am Chem Soc* 123(46): 11341-11352.

Wang, G., N. Sawai, S. Kotliarova, I. Kanazawa and N. Nukina (2000). Ataxin-3, the MJD1 gene product, interacts with the two human homologs of yeast DNA repair protein RAD23, HHR23A and HHR23B. *Hum Mol Genet* 9(12): 1795-1803.

Winborn, B. J., S. M. Travis, S. V. Todi, K. M. Scaglione, P. Xu, A. J. Williams, R. E. Cohen, J. Peng and H. L. Paulson (2008). The deubiquitinating enzyme ataxin-3, a polyglutamine disease protein, edits Lys63 linkages in mixed linkage ubiquitin chains. *J Biol Chem* 283(39): 26436-26443.

Wishart, D. S., B. D. Sykes and F. M. Richards (1992). The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry* 31(6): 1647-1651.

Yao, J., H. J. Dyson and P. E. Wright (1997). Chemical shift dispersion and secondary structure prediction in unfolded and partly folded proteins. *FEBS Lett* 419(2-3): 285-289.

Zhong, X. and R. N. Pittman (2006). Ataxin-3 binds VCP/p97 and regulates retrotranslocation of ERAD substrates. *Hum Mol Genet* 15(16): 2409-2420.

2.5 The conformational ensemble of the third Ubiquitin Interacting Motif of Ataxin-3.

The study has been carried out in collaboration with the Structural Biology and NMR laboratory group at the Department of Biology of the University of Copenhagen (Denmark) and the group of Prof. Gary S. Shaw at the Department of Biochemistry University of Western Ontario, London, Ontario, (Canada)

2.5.1 Introduction

Protein misfolding diseases are a class of pathologies caused by the partial loss of native structure of specific proteins leading to a loss of physiological function and usually to abnormal aggregation. Among these diseases the most studied and well known are amyloid diseases that share as hallmark the formation of large and highly stable intermolecular β-sheet fibrils as final product of their aggregation. Amyloid diseases are often associated to neurodegeneration, and comprise Alzheimer, Parkinson, Huntington, etc

A particular class of neurodegenerative amyloid diseases are the polyglutamine (polyQ) diseases. Proteins responsible of these pathologies have different structures and functions but all share as a distinctive feature a polyQ tract. PolyQ expansion over a certain threshold causes protein misfolding and aggregation leading to amyloid formation and neurodegeneration (Orr et al., 2007). This class of proteins does not share any functional and structural similarity besides the fact that they are multidomain proteins harbouring a polyQ tract. It has been largely demonstrated that although the expansion of the poly stretch is the causative agent of the neurodegenerative disorder, the amino acidic context in which the polyQ is located also plays a crucial role in the aggregation process.

Protein regions flanking the polyQ can thus have a great influence on the aggregation pathway and consequently on the toxicity of the formed abnormal species (Duennwald et al. 2006). In this context, the alternative splicing observed in many polyQ proteins is of particular interest as it could also influence the regions flanking the polyQ. Indeed, alternative splicing is an essential process to ensure high proteomic diversity and most of the human genes codifies for more than one transcript variant (Johnson et al., 2003, Wang et al., 2008).

Human ataxin-3 (AT3) is a multi-domain polyQ deubiquitinating enzyme of 43 kDa that is often used as a model system for the study of polyQ diseases and undergoes alternative splicing (Harris et al., 2010). AT3 is composed of three distinct domains; the structured N-terminal Josephin Domain (JD), a cysteine-protease domain, a mainly disordered region between the JD and the polyQ stretch (residues 182-291), which contains two ubiquitin (Ub) interacting motifs (UIMs, Fisher et al., 2013), and a C-terminal part which undergoes alternative splicing responsible for the main AT3 isoforms. Among the different AT3 isoforms, the first isolated contains a cluster of hydrophobic and aromatic residues in the last 50 residues of the C-terminal region whereas a second isoform presents an additional UIM replacing the hydrophobic tail (Figure 1) (Goto et al., 1997, Bettencourt et al., 2010). These alternative splicing isoforms have been identified as 2UIM and 3UIM, respectively. The 3UIM isoform is widely expressed and seems to be the predominant form in the human brain (Harris et al., 2010). Both 2UIM and 3UIM are functional proteins and the different C-terminal splicing does not alter the deubiquitinating activity of AT3 (Harris et al., 2010). It has been shown, however, that the alternative splicing at the C-terminal influences the tendency of AT3 to aggregate and affects its stability, with 2UIM more prone to aggregate and less stable than 3UIM. Further, the third UIM in the C-terminal region of AT3 is involved in a multivalent binding to Parkin ubiquitin ligase domain (Bai et al. 2013).

Nothing is known from the structural point of view about the two C-terminal regions of 2UIM and 3UIM AT3 isoforms. The knowledge of the ensemble of structures of different variants of a protein is fundamental for a better understanding of protein functional and misfunctional properties. Molecular dynamics (MD), in particular, is a powerful tool to characterize heterogeneous protein ensembles in atomistic details and can provide a high agreement with biophysical experimental data when accurate force fields and sufficient sampling of the conformational space is provided (Klepeis et al., 2009). Thus, we here report the first characterization of structure and dynamics of the 3UIM C-terminal regions of AT3 isoform using explicit solvent enhanced sampling all-atom simulations based on temperature replica exchange and two different state-of-the-art protein force fields (i.e., the physical models used to describe the protein and the solvent during the simulations), which have been recently applied to the study of partially disordered, heterogeneous and unfolded proteins (Knot and Best, 2012, Camilloni and Vendruscolo, 2014, Stanley et al., 2014). The results for 3UIM have been also compared to NMR data already available for other UIM motifs in solution. We also investigated, employing the same approach, the effects induced by mutations of the 3UIM variant designed to introduce five residues of the 2UIM variant that are predicted as hotspots of protein aggregation (E336P, D338F, Q341F, V344F, S347Y) along with a deletion (E337Δ).

2.5.2 Results and Discussion

REMD simulations to study the conformational ensemble of the C-terminal region of 3UIM

Here, we employed explicit solvent Replica-Exchange Molecular Dynamics (REMD) simulations to characterize the conformational ensemble in solution of the 3UIM C-terminal regions of AT3 (residue 306-361). We have selected the REMD approach since it has the potential to enhance the sampling in

MD simulations without renouncing to an atomistic description of both the protein and the solvent, as attested by its applications to many cases of study (Knott and Best, 2012). In particular, we used 64 replicas in our REMD calculations with temperatures spanning from 299 to 360 K, collecting overall more than 3.2 µs per system. We also evaluate the influence of different force-field descriptions for both the protein (Amber03w and CHARMM22*) and solvent models (TIP3P, TIPS3P, TIP4P2005) to assess the reproducibility of the results. This approach allows us *i)* to verify it the 3UIM C-terminal region is or is not stable in a helical conformation in solution *ii)* to estimate the population of the helical conformations and compare it to the available information on other UIM motifs in solution *iii)* to identify in the ensemble of the free fragment in solution the existence of conformations that are Ub-bound like comparing our ensemble to the experimentally known Ub-bound UIM structures of other proteins *iv)* to evaluate if the introduction of the aggregation-prone residues of 2UIM in the 3UIM sequence modulates the conformational ensemble of the protein and their impact on the helical structure.

Helical structures of 3UIM are not stable in the free state of the domain in solution.

UIMs are thought to assume an α-helical structure also in the free state (Hofmann et al., 2001), when Ub is not bound. Nevertheless, most of UIM structures have been solved by X-ray crystallography and in complex with other biological partners. It is not clear if a UIM motif populates in the free ensemble in solution only helical conformations. In the C-terminal region of 3UIM AT3, the UIM spans the residues E336-T350 (Donaldson et al., 2003). This region is predicted to be helical by PSI-PRED for 3UIM AT3 (S335-K356) (Figure 1). We thus modelled this region as a α-helix in the starting structures for the simulations and we then monitored its evolution over the simulation time. We selected the replica at 304 K as a reference for our analysis since it was previously used in other works that applied a similar approach for the study of disordered domain (Knott and Best, 2012) (Bai et al., 2013). We first calculated the helical content from each 304 K replica. The UIM region of 3UIM has generally helical content higher than 20% despite the changes in the force fields or water models (Figure 2). We observed, however, that when the A03w-TIP4P2005 force field was employed the helical content for the same region is even higher (> 40%). In our ensemble the UIM residues do not always consistently populate a stable helix. Indeed, we observed a breaking of the helix in correspondence of the residues Ala343-Ser347, and the 3UIM fragment also populates disordered conformations, which are a major population of the ensemble when CHARMM22* was used. Moreover, all the simulations point out that a sub-region of the UIM, from residues Glu337 to Val344, presents higher helical propensity around the 30%.

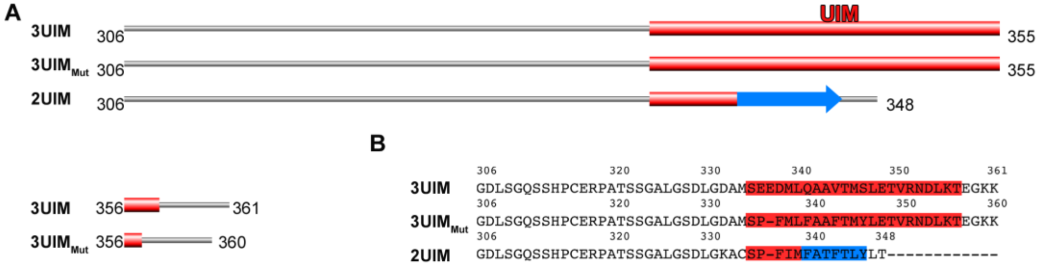

A
UIM
3UIM 306
355
3UIM$_{Mut}$ 306
355
2UIM 306
348
3UIM 356
361
3UIM$_{Mut}$ 356
360
B
3UIM 306 320 330 340 350 361
GDLSGQSSHPCERPATSSGALGSDLGDAMSEEDMLQAAVTMSLETVRNDLKTEGKK
3UIM$_{Mut}$ 306 320 330 340 350 360
GDLSGQSSHPCERPATSSGALGSDLGDAMSP-FMLFAAFTMYLETVRNDLKTEGKK
2UIM 306 320 330 340 348
GDLSGQSSHPCERPATSSGALGSDLGKACSP-FIMFATFTLYLT------------

Figure 1. Secondary structure prediction of the AT3 2UIM, 3UIM and 3UIM$_{Mut}$ isoforms. A)The prediction was performed using PSI-PRED. Regions predicted as α-helix are represented as red box while the ones predicted as β-strand are represented as blue arrow. B) The predicted secondary structure elements are represented on the aminoacidic sequence of each isoforms.

To discriminate if the low stability of a helical conformation in solution is a distinctive trait of 3UIM or a more common property of other UIM motifs, we have searched in the NMR database BMRB for chemical shifts data on UIM motifs in solution. We have identified 4 structures with released chemical shifts, 3LDZ, 2KDI, 1X2Q and 2MKG (which comprises two UIM motifs in its sequence). We then predicted by δ2D (Camilloni et al., 2012) the secondary structure propensity from the chemical shifts. We observed a high variability of helical content in the UIMs investigated, ranging from structures with high a-helix content (around 0.8 for both UIMs from 2MKG, 3LDZ) to structures with a rather low propensity comparable with those observed from our CHARMM22* simulations (lower than 0.5 for 2KDI and 1X2Q) (Figure 3). We have then compared 3UIM primary sequence with other known UIMs and with the consensus sequence deposited in the PFAM database (PF02809) and consensus sequence obtained from literature (Sgourakis, 2010) (Figure 3). The analysis shows that the UIM in the C-terminal region of 3UIM isoform presents the typical features of a UIM motif with highly conserved residues, such as Leu340, Ala343, Ser346, acidic residues in the N-terminal part of the motif (Glu336-Asp338) and the pattern of hydrophobic residues (Figure 3). Moreover, we should notice that suboptimal residues are present in the 3UIM sequence in comparison to the other UIMs, like Val344 and Thr345, that have low helix propensity (Pace and Scholtz, 1998, Best et al., 2012) and they are localized in the region of 3UIM where the helix breaks in some of the simulation frames (see above).

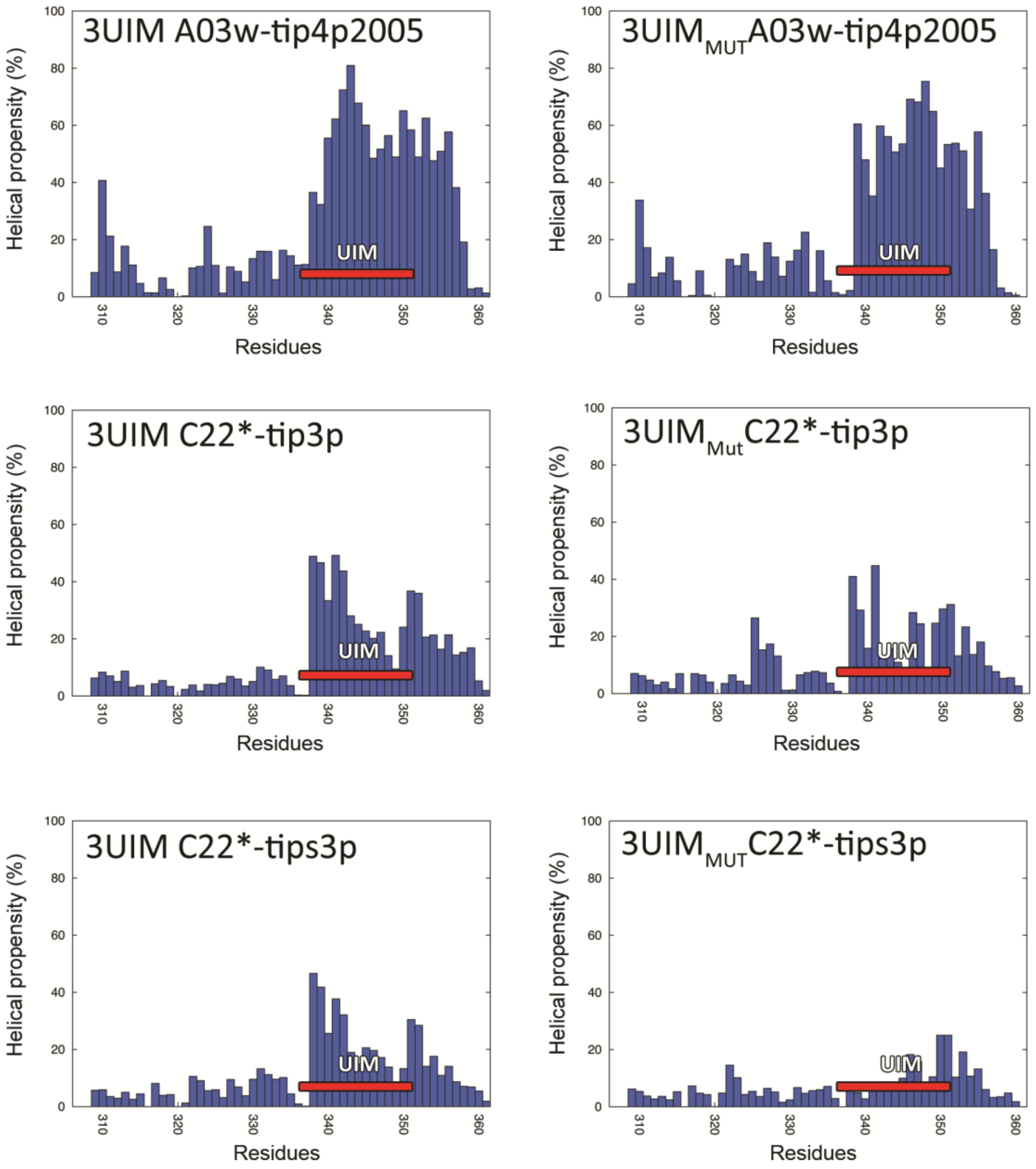
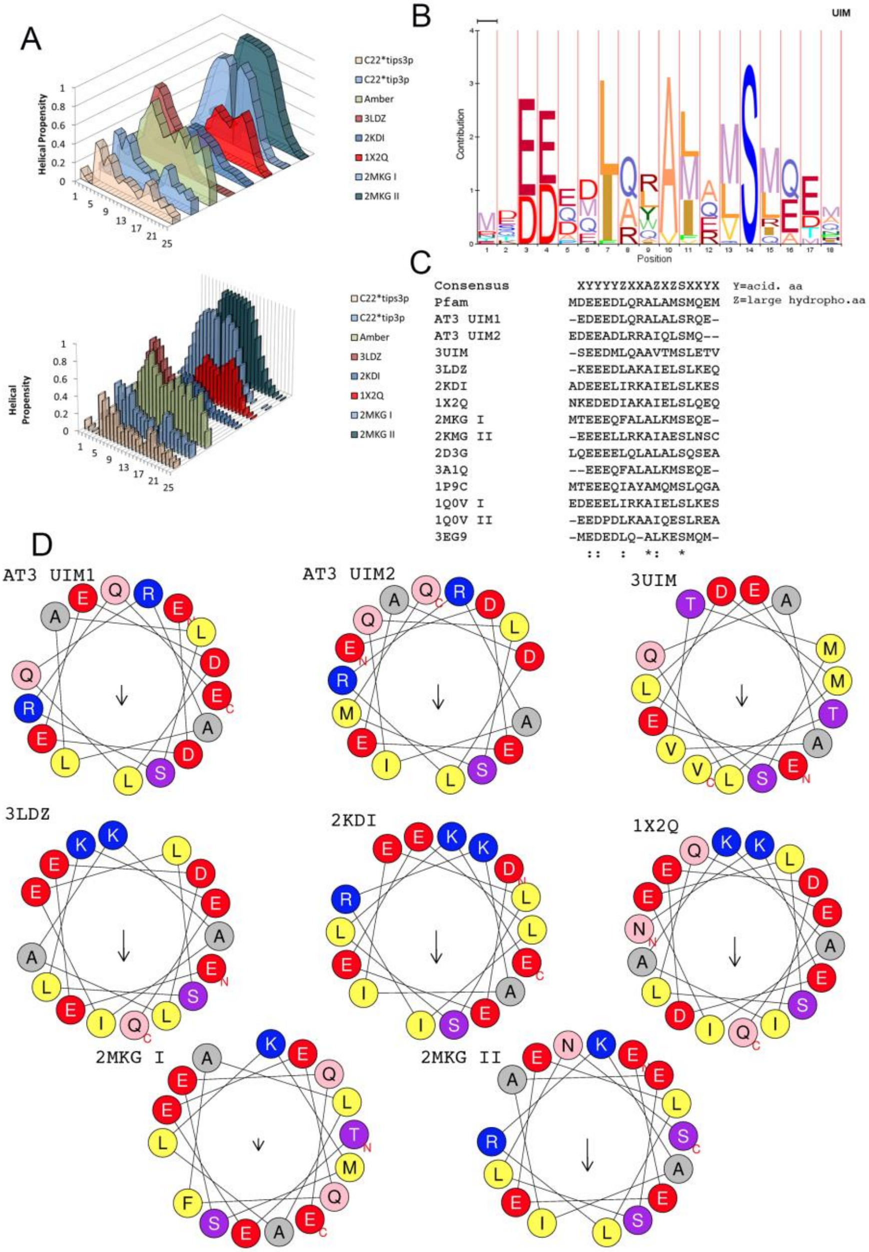


Figure 2. Helical content at 304 K. The per-residue helical content of the replicas at 304 K from the REMD simulations of 3UIM and $3UIM_{Mut}$ are shown for each combination of protein force field (Amber03W, CHARMM22*) and water model (TIP4P2005, TIP3P, TIPS3P). The residues belonging to the UIM (residues 336-350) of the 3UIM AT3 variant are highlighted with a red bar.



```
Consensus      XYYYYZXXAZXZSXXYX  Y=acid. aa
Pfam          MDEEEDLQRALAMSMQEM  Z=large hydropho.aa
AT3 UIM1      -EDEEDLQRALALSRQE-
AT3 UIM2      EDEEADLRRAIQLSMQ--
3UIM          -SEEDMLQAAVTMSLETV
3LDZ          -KEEEDLAKAIELSLKEQ
2KDI          ADEEELIRKAIELSLKES
1X2Q          NKEDEDIAKAIELSLQEQ
2MKG I        MTEEEQFALALKMSEQE-
2KMG II       -EEEELLRKAIAESLNSC
2D3G          LQEEEELQLALALSQSEA
3A1Q          --EEEQFALALKMSEQE-
1P9C          MTEEEQIAYAMQMSLQGA
1Q0V I        EDEEELIRKAIELSLKES
1Q0V II       -EEDPDLKAAIQESLREA
3EG9          -MEDEDLQ-ALKESMQM-
                :: :  *:  *
```

Figure 3. Structural properties of UIMs. A) Secondary structure propensity predicted from chemical shifts by δ2D. B) Consensus sequence for the UIM domain in the PFAM database. C) Alignement of the primary sequences of UIM motifs identified. D) Helical wheel representation of the UIMs investigated.

We also provided a helical wheel representation of the AT3 UIM1, UIM2, 3UIM and of the UIMs previously investigated (3LDZ, 2KDI, 1X2Q and 2MKG I and II) (Figure 3). The analysis show that when the 3UIM assumes a helical conformation the Thr345 is located under one of the acidic residues, Glu338, in the N-terminal part, that are highly conserved in the UIMs. Moreover the Thr350 is located on the same face of the helix and under the Ala343, a residue that is strictly conserved in all the UIMs since it is involved in the interaction with ubiquitin (Fisher et al., 2003). We suggest that the location of suboptimal residues, especially threonine, in the 3UIM have a role into its low propensity to populate stable helical conformations in solution.

The UIM motif of 3UIM is replaced by an aggregation-prone region in 2UIM variant of AT3

Since two major isoforms are known for AT3 and the differences are located in the C-terminal region and affect aggregation, we also predicted the aggregation-prone regions (APRs) of the C-terminal regions of both 3UIM and 2UIM isoforms using a consensus across five different programs: Aggrescan (Vendrelli et al., 2007), Tango (Linding et al., 2004), Zyggregator (Tartaglia and Vendruscolo, 2008), Waltz (Maurer-Stroh et al., 2010) and Pasta (Walsh et al., 2014). All the predictors highlight the presence of an APR in the 2UIM variant from residues 337-348, which is lost in the 3UIM variant. The different aggregation propensity is related to the substitution of charged (E336, E337, D338, Q341 and S347) with hydrophobic or aromatic residues (P336, F337 F340, F343 and Y346). As expected, a $3UIM_{Mut}$ variant where the 3UIM residues in the region corresponding to 2UIM APR are mutated to hydrophobic and aromatic residues found in the 2UIM isoform (E336P, E337Δ, D338F, Q341F, V344F and S347Y) exhibits a higher aggregation propensity (Figure 4). One should also notice that this region overlaps with the UIM of 3UIM isoform. We thus decided to apply the REMD approach to the $3UIM_{Mut}$ variant to predict the effects that these mutations have on the conformation of the fragment in solution.

The replacement of residues in the UIM motif of 3UIM with 2UIM aggregation-prone residues alter the structural ensemble of 3UIM fragment

The mutations of 3UIM with aggregation-prone residues of 2UIM variant decrease the helical propensity by more than three-two folds, indicating that these few substitutions not only change the aggregation propensity (Harris et al., 2010) of the protein but also its local structure. In this context, CHARMM22*/TIPS3P is the more effective force field in describing the loss of helical content upon mutations, whereas with Amber03w-TIP4P2005 is less evident. As stated before, all the simulations of

3UIM point out that a sub-region of the UIM, from residues Glu337 to Val344, presents higher helical propensity, more than 40% that is likely to be a fully formed helix in the sampled ensemble. This sub-region comprises two negatively charged residues (Glu337 and Asp338) and an aliphatic cluster (Ala342, Ala343, Val344) in 3UIM that we mutated to the corresponding residues in 2UIM. In the $3UIM_{mut}$ the presence of these mutations abolishes the helical propensity in that sub-region of the UIM, by an average of 35%, suggesting the importance of these residues in modulating its structural properties.

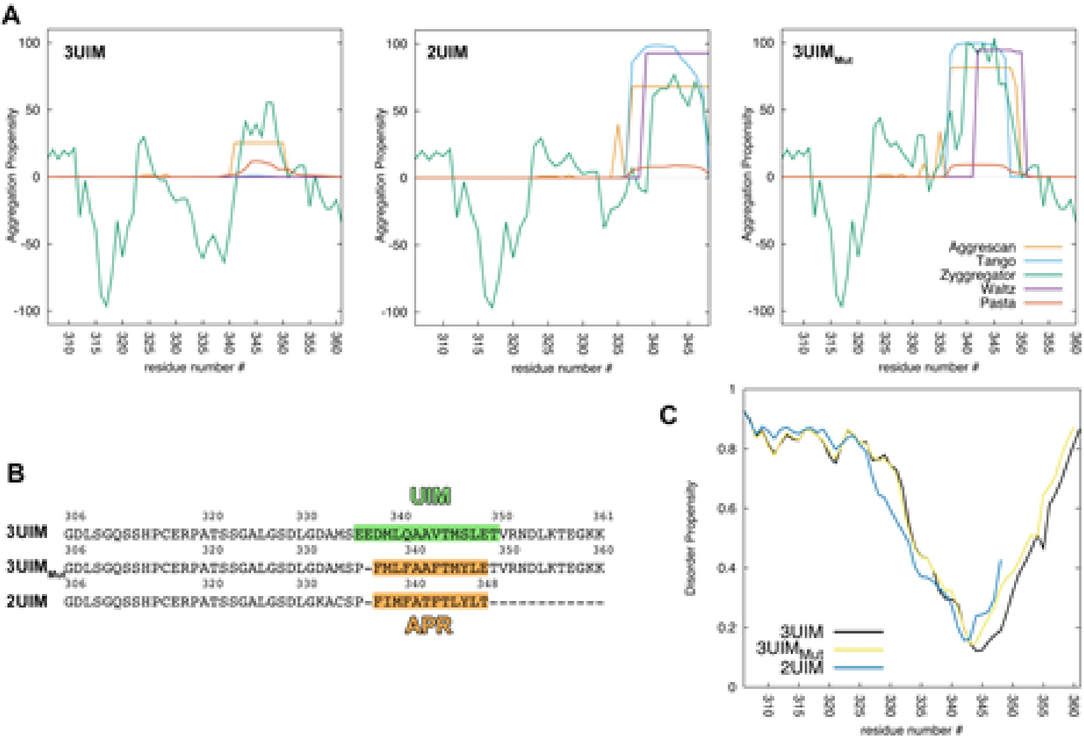


Figure 4 Prediction of the aggregation-prone regions (APRs) of the C-terminal portion of Ataxin-3 2UIM and 3UIM isoforms. We analyzed 2UIM, 3UIM and a mutant variant of 3UIM ($3UIM_{Mut}$ E336P, D338F, Q341F, V344F, S347Y, E337Δ). A) The APR prediction was carried out using five different programs (Aggrescan, Tango, Zyggregator, Waltz, Pasta). B) The sequences of each of the three constructs considered are showed along with the APR region predicted for the 2UIM and $3UIM_{Mut}$ isoforms and the helical region predicted for the 3UIM by PsiPred. C) Plot of residue-dependent disorder propensity prediction carried out using three different programs (Disoclust, Disopred, PreDisorder) and the average consensus between their results is showed for each isoforms.

To further characterize the REMD ensemble of 3UIM and 3UIM$_{mut}$, we calculated the Rg of each 304K replica. The results are shown in the two-dimensional (2D) probability density plot against the helical propensity calculated for only the UIM region (Figure 5). The probability density function is represented as a contour plot on which the blue area corresponds to lower probability density while the red region can be identified as clusters of largely populated states. The REMD ensembles that we collected are very sensitive to the different protein force fields and waters models employed in the simulations with regard to compactness of the protein both for 3UIM and 3UIM$_{Mut}$ variants. The Rg values show that the mutations in the 3UIM$_{mut}$ alters the structural properties of 3UIM towards more compact structures, from an average Rg of 1.45 ± 0.24 nm to 1.27 ± 0.19 nm in the simulations with CHARMM22*/TIPS3P.The simulations with CHARMM22*/TIP3P sample less extended states both for the 3UIM and 3UIM$_{mut}$ (avg. 1.36 ± 0.27 and 1.27 ± 0.19 nm), than the other two simulations settings. On the contrary in 3UIM simulations performed with the model of water to TIPS3P and CHARMM22* there is a clear less propensity to populate compact states towards more open ones (average Rg 1.45 ± 0.24 nm). The 3UIM simulations with Amberff03w and TIP4P2005 have average Rg values close to the CHARMM22*/TIPS3P, but it samples more often conformations more extended (with Rg values till 2.8 nm).

The results that we collected so far show that five mutations and one deletion introduced in 3UIM construct are sufficient to modulate the conformational ensemble, destabilizing the helical ubiquitin interacting motif and shifting the ensemble toward more disordered structures. The combination of different protein force fields and water models allowed us to achieve a more complete picture of the effects induced by the mutations. Overall, the different force fields provide a similar description of the changes induced on the secondary and tertiary structure, whereas if we look at the data at a finer level some differences can be identified. The results thus suggest that when we compare protein variants in a MD framework, we can benefit from the usage of several distinct force fields to avoid over-interpretation of the results.

Ubiquitin-binding conformations of the AT3 3UIM

We know that many free proteins in solution can sample bound-like states (Baldwin and Kay, 2009). We thus wonder if this is the case of the UIM region of 3UIM. We thus compared the structures of our 3UIM REMD ensembles with the experimental structure of the Ub monomer in complex with the UIM region of the yeast vps27 protein (PDB ID 1Q0V) and the proteasome subunit S5a (PDB ID 1YX5). We analysed each REMD ensemble (replicas at 304 K) of 3UIM in 10 structural clusters using the Gromos algorithm and a root-mean square deviation (RMSD) cutoff of 1 nm.

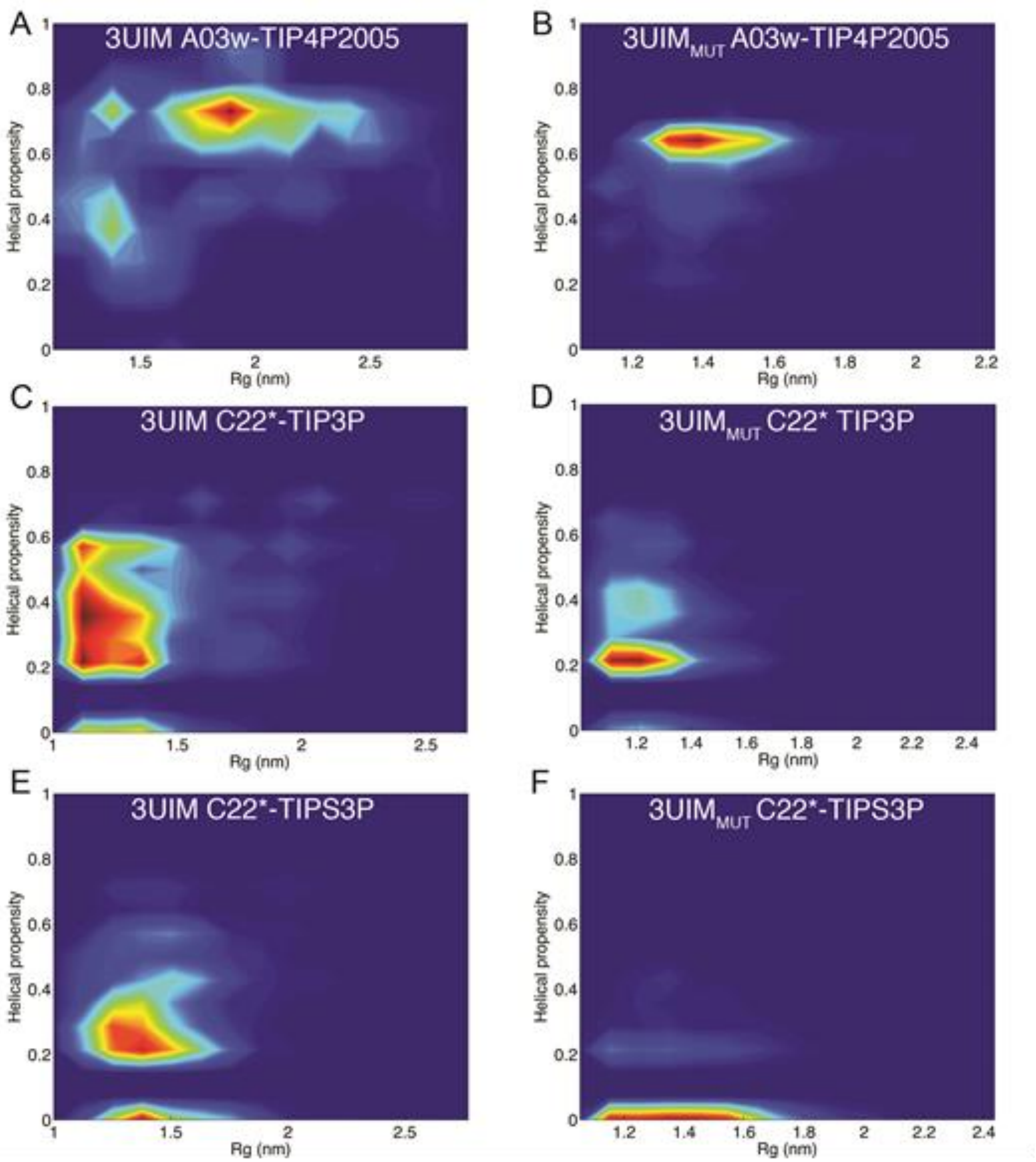


Figure 5. Radius of gyration (Rg) and helical content calculated for only the UIM region. The two-dimensional probability density distribution of Rg and SAS of replicas at 304 K of each system is here shown as a contour plot from blue (low populated regions) to red (highly populated regions).

We then superimpose the average structure of each cluster with the UIMs of vps27-Ub and S5a-Ub complexes. We have been able to identify 3UIM conformations, in the main clusters, characterized by low RMSD (lower than 1 Å) with respect to the experimental complexes with all the combinations of protein force fields and water models (Figure 6).

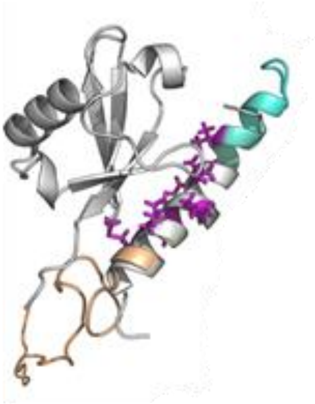


Figure 6. Conformations similar to the structure of other UIMs in complex with the ubiquitin are found in the free ensemble of 3UIM C-terminal region of AT3 in solution. The conformation from 3UIM CHARMM22*/TIPS3P replicate at 304 K is highlighted with shade of color from N-terminal (wheat color) to C-terminal (cyan) superimposed to the structure of vsp27 UIM (white PDB ID 1Q0V).

In all these structure, the UIM sequence is a well-folded α-helix, as the canonical UIMs. We estimated the presence of bound-like α-helix on REMD ensembles by RMSD analysis of the 3UIM (residues 336-350) using as reference the structure of the first UIM of vps27 in complex with Ub (residue 259-273, PDB ID 1Q0V). We used a cutoff of 1.5 Å of RMSD to identify Ub bound-like states, since it well discriminates between well-folded α-helix and distorted conformations (Figure 7). In the 3UIM simulations performed with the model of water TIP3P or TIPS3P and CHARMM22* Ub bound like conformations are populated for the 7% and 5% of the REMD ensemble, respectively. From the comparison of these structures of 3UIM and Ub-bound UIMs of vps27 and S5a we have been able to identify a group of 3UIM residues that are predicted to interact with Ub (i.e., Glu336, Glu337, Met339, Leu340, Ala342, Ala343, Val344, Met346, Ser347, Leu348, Val351). The hydrophobic residues in this group are placed at favourable positions to form a solvent-exposed interface that may fit the hydrophobic binding site of Ub, that is located at the C-terminal three strands of the Ub β-sheet and include Leu8, Leu43, Ile44, Leu50, Leu69, and Leu71. Moreover the negatively charged residue are suggested in Ub-UIMs complexes to interact with three basic residues on Ub, Arg42, Lys48, Arg72 that are located near the hydrophobic binding site. In the $3UIM_{mut}$ REMD ensembles, we identify a lower propensity toward α-helical structures and conformations similar to the experimental Ub-bound structures are just one minor fraction of the ensemble (0.3% for CHARMM22*/TIP3P and 0% for CHARMM22*/TIPS3P simulations).

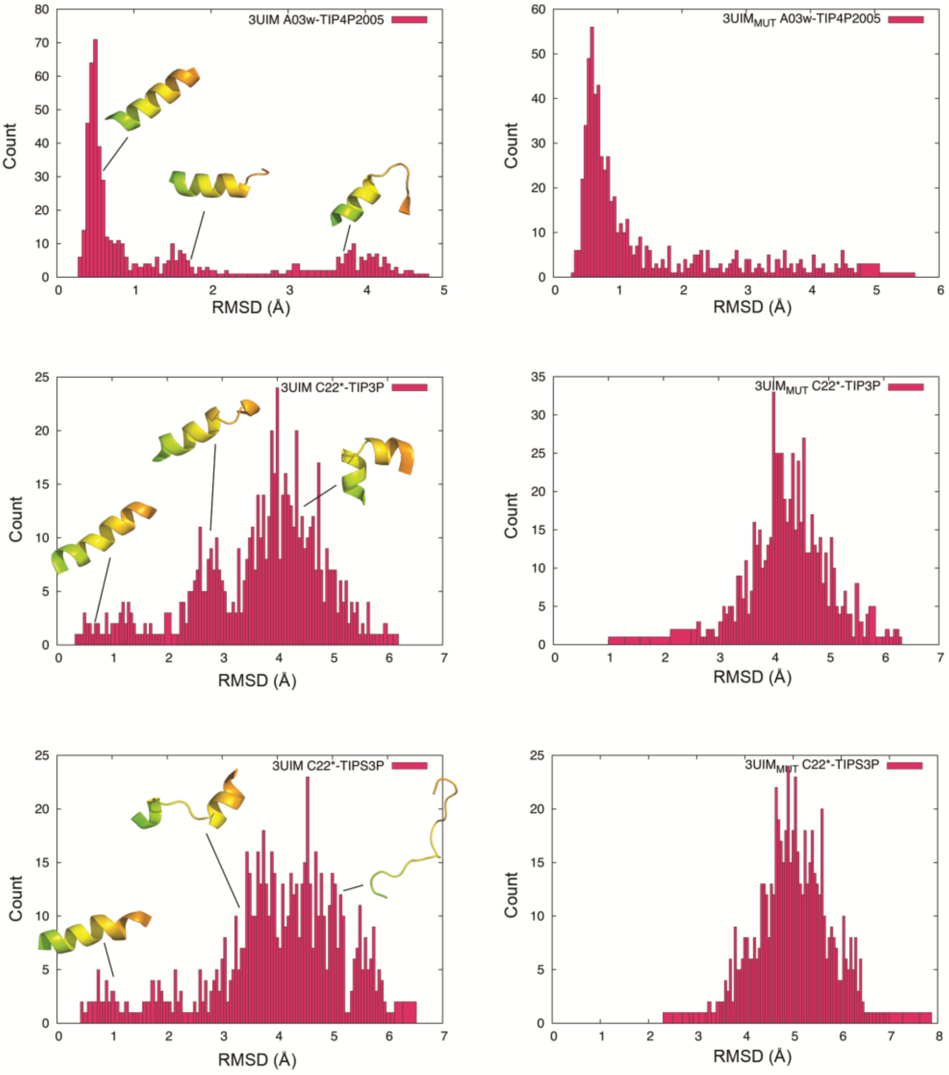


Figure 7. RMSD analysis of the AT3 3UIM (residues 336-350) using as reference the structure of the first UIM of vps27 in complex with ubiquitin (PDB ID 1Q0V residue 259-273). Structures of AT3 3UIM are shown as cartoon.

In summary, our comparison of our REMD ensembles with experimentally-determined Ub-UIM complexes suggest that the AT3 C-terminal UIM free in solution can be characterized by helical conformations competent for binding with Ub but they are very transient during our simulations.

2.5.3 Materials and Methods

Replica-Exchange Molecular Dynamics simulations

Replica-Exchange Molecular Dynamics (REMD) simulations of the free state of the wild type AT3 C-terminal region of 3UIM isoform (56 residues, 306-361) and a 3UIM mutant variant (E336P, E337Δ, D338F, Q341F, V344F and S347Y, 3UIM$_{Mut}$, 55 residues). In the temperature Replica Exchange scheme (Sugita and Okamoto, 1999), a number of different copies (replicas) of the system are simulated in parallel at different temperatures and exchanges of configurations are tried periodically between pairs of replicas. The advantage of this method is that if the trajectory is temporarily trapped in a local minimum can exchange with a higher-temperature replica and it can thus, more easily, cross high-energy barriers and allow the system to reach equilibrium more quickly.

We started from a model of AT3 3UIM and 3UIM$_{mut}$ generated with Crystallography and NMR System version 1.3 (CNS, Burger, 2007) on which we further imposed a helical structure according to the prediction by PSI-Pred. using MODELLER9.14 (Eswar et al., 2006).. We selected as starting structures for MD simulations, the models that lack side chain–side chain long-range (defined as contacts between residues at a distance in the sequence higher than 3) intramolecular contacts. The models were then soaked in a dodecahedric box of water molecules with periodic boundary conditions and the box was built so that all the protein atoms were at a distance of at least 1.2 nm from the box edges. We employed two different protein force fields and three water models in our simulations: i) the Amber ff03w force field (Best et al., 2010) was used for the protein (this force field has been adapted for the use with the TIP4P/2005 water model) with the water model TIP4P/2005 (Vega et al., 2005). ii) the CHARMM22* force field (Piana et al., 2011) was used with either the TIP3P or TIPS3P water models (Jorgensen et al. 1983). REMD simulations were performed by Gromacs 4.6.5 (www.gromacs.org), implemented on a parallel architecture. The LINCS algorithm was employed to constrain heavy atom bond lengths, allowing for the use of a 2 fs time-step. Long-range electrostatic interactions were calculated using Particle-Mesh Ewald (PME) method with a 0.12 nm grid spacing. Van der Waals and Coulomb interactions were truncated at 1.2 nm. Na+ and Cl− counterions were added to the system to

neutralize the overall charge and to simulate a physiological ionic strength (150 mM) according to protocols previously employed (Arrigoni et al., 2012, Lambrughi et al., 2012, Invernizzi et al., 2013).
Each system was initially relaxed by 10000 steps of energy minimization by the steepest descent method. The optimization step was followed by 50 ps of solvent equilibration at 300K, while restraining the protein atomic positions using a harmonic potential. The systems were subsequently simulated for 5 ns at 300 K at a constant pressure of 1 bar (NPT ensemble) using Nosé-Hoover thermostat and Parrinello-Rahman barostat with coupling constants of 5 and 10 ps respectively. From these NPT trajectories, a conformation was selected which had a volume close to the average volume of the trajectory and used as the starting point for the subsequent NVT preparatory step at 300K and the same thermostat for 20 ns. The 64 initial conformations for the REMD simulations were then selected from different points (between 10 and 20 ns) along this NVT trajectory.
We then carried out 50-ns REMD simulations using 64 replicas each running at a different temperature in the range between 299 K and 360 K. The temperature spacing between each neighboring replica has been selected to ensure an exchange probability higher than 0.2 between the neighboring replicas. Replicas exchanges were attempted every 10 ps.

Analyses of the simulations

The 304 K replica was used as a reference for the analysis, as done in other works (Knott and Best, 2012) and to compare with already available NMR data (Bai et al., 2013). Only to study the temperature distributions we converted each replica to be continuous to the simulation time in order to follow each replica through the temperature space. The frequency of the system’s visits to each temperature for each replica were used to check the convergence of the results. The secondary structure (ss) content was calculated by g_helix Gromacs tool, along with a residue-dependent helicity persistence profile. The pairwise main chain root mean square deviation (rmsd) matrix was calculated for each REMD ensemble. The Gromos algorithm was employed for clustering, using a cutoff of 1.1 nm calculated from the average values of rmsd matrix. For each cluster, the structure with the lowest rmsd compared to the other cluster members was selected as the average structure.

2.5.4 Conclusions

The UIMs are a class of ubiquitin-binding domain involved in a broad range of regulatory cellular mechanisms that are ubiquitination-mediated, such as protein degradation, quality control, endocytosis, cell-cycle control, DNA repair, signalling and transcription (Hofmann and Falquet, 2001). UIMs often promote autoubiquitination, regulating the activity of the protein that contains them and they are α-helical and interact with a highly conserved alanine residue to Ile44 of the Ub hydrophobic patch. AT3 exists in numerous isoforms and the first isolated contains a cluster of hydrophobic and aromatic residues in the last 50 residues of the C-terminal region whereas a second isoform presents an additional UIM replacing the hydrophobic tail (Goto et al., 1997, Bettencourt et al., 2010, Harris et al., 2010). This region is located near the polyQ tract and it has been pointed out that region flanking polyQ can have important role in modulate the aggregation and in disease. Moreover the knowledge of the ensemble of structures of different variants of a protein is fundamental for a better understanding of protein functional and misfunctional properties. Since nothing is known from the structural point of view about this region of AT3, we investigated the C-terminal region of AT3 3UIM isoform employing Replica-Exchange MD (REMD) simulations and two different force fields and two different solvent models. The results obtained with the two force fields are not in agreement and show relevant differences in the helical propensity and in the compactness of AT3 3UIM. We compared our results with NMR data available (Bai et al., 2013), concluding that the simulations performed with CHARMM22* force field and the TIPS3P water model give a reasonable description helical content for AT3 3UIM, but unrealistic overcollapsed structures. These results provide one more example of the limitation of current MD force fields in describing IPDs conformational states, a problem which can be mitigated by the development of *ad-hoc* force fields for IDPs. Despite these limitations, we provided a first description of the conformational states of the third UIM of AT3 showing that it has a low propensity to populate helical conformations in solution. We analysed the structural properties of C-terminal region of AT3 3UIM and provided results to understand its low helical propensity. We suggest that the location of suboptimal residues for the geometry of an α helix in the 3UIM have a role into its low propensity to populate stable helical conformations in solution. We also identified, in 3UIM free ensemble, conformations that resemble the NMR structure of a canonical UIM is complex with ubiquitin. Due to the low propensity for helical states in solution, further studies are necessary to clarify if 3UIM is not a canonical UIM and thus bind with very low affinity or cannot bind ubiquitin, or if a disordered ensemble can be a common feature of also other UIM domains.

2.5.5 References

Arrigoni, A., B. Grillo, A. Vitriolo, L. De Gioia and E. Papaleo (2012). C-terminal acidic domain of ubiquitin-conjugating enzymes: A multi-functional conserved intrinsically disordered domain in family 3 of E2 enzymes. *Journal of Structural Biology* 178(3): 245-259.

Bai, J.J., S.S. Safadi, P., Mercier, K.R. Barber, G.S. Shaw (2013). Ataxin-3 is a multivalent ligand for the parkin Ubl domain. *Biochemistry* 52(42):7369-76.

Baldwin, A. J., and L. E., Kay (2009). NMR spectroscopy brings invisible protein states into focus. *Nat. Chem. Biol.* 5, 808–14.

Best, R. B., D. de Sancho and J. Mittal (2012). Residue-specific alpha-helix propensities from molecular simulation. *Biophys J* 102(6): 1462-1467.

Best, R. B. and J. Mittal (2010). Protein simulations with an optimized water model: cooperative helix formation and temperature-induced unfolded state collapse. *J Phys Chem B* 114(46): 14916-14923.

Bettencourt, C., C. Santos, R. Montiel, C. Costa Mdo, P. Cruz-Morales, L. R. Santos, N. Simoes, T. Kay, J. Vasconcelos, P. Maciel and M. Lima (2010). Increased transcript diversity: novel splicing variants of Machado-Joseph disease gene (ATXN3). *Neurogenetics* 11(2): 193-202.

Brunger, A. T. (2007). Version 1.2 of the Crystallography and NMR system. *Nat Protoc* 2(11): 2728-2733.

Camilloni, C., A. De Simone, W. F. Vranken and M. Vendruscolo (2012). Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. *Biochemistry* 51(11): 2224-2231

Camilloni, C., P. Robustelli, A. De Simone, A. Cavalli and M. Vendruscolo (2012). Characterization of the conformational equilibrium between the two major substates of RNase A using NMR chemical shifts. *J Am Chem Soc* 134(9): 3968-3971.

Camilloni, C. and M. Vendruscolo (2014). Statistical mechanics of the denatured state of a protein using replica-averaged metadynamics. *J Am Chem Soc* 136(25): 8982-8991.

Conchillo-Solé, O., S. N, de Groot, F. X., Avilés, J., Vendrell, X., Daura, S., Ventura (2007) AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics* 2007, 8:65.

Donaldson, K. M., W. Li, K. A. Ching, S. Batalov, C. C. Tsai and C. A. Joazeiro (2003). Ubiquitin-mediated sequestration of normal cellular proteins into polyglutamine aggregates. *Proc Natl Acad Sci U S A* 100(15): 8892-8897.

Duennwald, M. L., S. Jagadish, P. J. Muchowski and S. Lindquist (2006). Flanking sequences profoundly alter polyglutamine toxicity in yeast. Proc Natl Acad Sci U S A 103(29): 11045-11050.

Eswar, N., B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M. Y. Shen, U. Pieper and A. Sali (2006). Comparative protein structure modeling using Modeller. *Curr Protoc Bioinformatics* Chapter 5: Unit 5 6.

Fisher, R. D., B. Wang, S. L. Alam, D. S. Higginson, H. Robinson, W. I. Sundquist and C. P. Hill (2003). Structure and ubiquitin binding of the ubiquitin-interacting motif. *J Biol Chem* 278(31): 28976-28984.

Goto, J., M. Watanabe, Y. Ichikawa, S. B. Yee, N. Ihara, K. Endo, S. Igarashi, Y. Takiyama, C. Gaspar, P. Maciel, S. Tsuji, G. A. Rouleau and I. Kanazawa (1997). Machado-Joseph disease gene products carrying different carboxyl termini. *Neurosci Res* 28(4): 373-377.

Harris, G. M., K. Dodelzon, L. Gong, P. Gonzalez-Alegre and H. L. Paulson (2010). Splice isoforms of the polyglutamine disease protein ataxin-3 exhibit similar enzymatic yet different aggregation properties. *PLoS One* 5(10): e13695.

Hofmann, K. and L. Falquet (2001). A ubiquitin-interacting motif conserved in components of the proteasomal and lysosomal protein degradation systems. *Trends Biochem Sci* 26(6): 347-350.

Invernizzi, G., M. Lambrughi, M. E. Regonesi, P. Tortora and E. Papaleo (2013). The conformational ensemble of the disordered and aggregation-protective 182-291 region of ataxin-3. *Biochim Biophys Acta* 1830(11): 5236-5247.

Johnson, J.M., J., Castle, P., Garrett-Engele, Z., Kan, P.M., Loerch, C.D., Armour, R., Santos, E.E., Schadt, R., Stoughton, R. Shoemaker (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science.* 302(5653):2141-4.

Jorgensen, W. L., J., Chandrasekhar, J.D., Madura, R. W. Impey, M. L. Klein, Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys* 1983, 79, 926-935.

Klepeis, J. L., K., Lindorff-Larsen, R. O., Dror, D. E., Shaw (2009). Long-timescale molecular dynamics simulations of protein structure and function. Curr Opin Struct Biol 19(2): 120-127.

Knott, M. and R. B. Best (2012). A preformed binding interface in the unbound ensemble of an intrinsically disordered protein: evidence from molecular simulations. *PLoS Comput Biol* 8(7): e1002605.

Lambrughi, M., E. Papaleo, L. Testa, S. Brocca, L. De Gioia and R. Grandori (2012). Intramolecular interactions stabilizing compact conformations of the intrinsically disordered kinase-inhibitor domain of Sic1: a molecular dynamics investigation. *Front Physiol* 3: 435.

Linding R., J., Schymkowitz,F., Rousseau, F., Diella, L., Serrano (2004) A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins. *J Mol Biol* 345-353, 2004

Lindorff-Larsen, K., S. Kristjansdottir, K. Teilum, W. Fieber, C. M. Dobson, F. M. Poulsen and M. Vendruscolo (2004). Determination of an ensemble of structures representing the denatured state of the bovine acyl-coenzyme a binding protein. *J Am Chem Soc* 126(10): 3291-3299.

Lindorff-Larsen, K., N. Trbovic, P. Maragakis, S. Piana and D. E. Shaw (2012). Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *J Am Chem Soc* 134(8): 3787-3791.

Maurer-Stroh, S., M., Debulpaep, N., Kuemmerer, M., Lopez de la Paz, IC, Martins, J., Reumers, K.L., Morris, A., Copland, L., Serpell, L., Serrano, J.W., Schymkowitz, F., Rousseau (2010). Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat Methods.*7(3):237-42.

Orr, H. T. and H. Y. Zoghbi (2007). Trinucleotide repeat disorders. *Annu Rev Neurosci* 30: 575-621.

Pace, C. N. and J. M. Scholtz (1998). A helix propensity scale based on experimental studies of peptides and proteins. *Biophys J* 75(1): 422-427.

Piana, S., K. Lindorff-Larsen and D. E. Shaw (2011). How robust are protein folding simulations with respect to force field parameterization? *Biophys J* 100(9): L47-49.

Sgourakis, N. G., M. M. Patel, A. E. Garcia, G. I. Makhatadze and S. A. McCallum (2010). Conformational dynamics and structural plasticity play critical roles in the ubiquitin recognition of a UIM domain. *J Mol Biol* 396(4): 1128-1144.

Stanley, N., S., Esteban-Martín , G., De Fabritiis (2014). Kinetic modulation of a disordered protein domain by phosphorylation. *Nature Communications* 5: 5272

Sugita, Y. and Y. Okamoto (1999). Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* 314(1–2): 141-151.

Tartaglia G., M., Vendruscolo (2008). The Zyggregator method for predicting protein aggregation propensities. *Chem Soc Rev.* 2008 Jul;37(7):1395-401.

Vega, C. and J. L. Abascal (2005). Relation between the melting temperature and the temperature of maximum density for the most common models of water. *J Chem Phys* 123(14):144504.

Walsh, I., F., Seno, S., Tosatto, A. Trovato (2014). PASTA 2.0: an improved server for protein aggregation prediction. *Nucl. Acids Res.* doi: 10.1093/nar/gku399.

Wang, E.T., R., Sandberg, S., Luo, I., Khrebtukova, L., Zhang, C., Mayr, S.F., Kingsmore, G.P., Schroth, C.B, Burge (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456(7221):470-6.

3 Computational characterization of the role of metal ions in the dynamical and functional properties of Zn^{2+} binding proteins: zinc-fingers and p53.

Metal ions are essential for the growth of all life forms (Bertini et al. 2001) and more than 40% of known proteins in eukaryote require the interaction with metal ions or metal containing cofactors to exploit their biological function (Dudev and Lim 2008). Metals help in protein folding or work within folded proteins to provide them functional properties and enhance or modulate them. Organisms exploit metal cofactors in proteins thanks to their unique features such as their small size and simple structure, high mobility and bioavailability, electron acceptor ability, positive charge, variable coordination sphere, specific ligand affinity, valence state and electron spin configuration. Thanks to these unique properties binding of metals ions is a common and widespread evolutionary strategy in protein, exploited to perform a wide range of different key biological processes such as protein structure stabilization, enzyme catalysis, cellular signaling and signal transduction, nitrogen fixation, photosynthesis, respiration and control of oxidative stress (Bertini et al., 2001). Metal ions are an integral part of many enzymes and are indispensable in several catalytic reactions, e.g., hydrolytic, redox and isomerization reactions. In particular, transition metals, such as Fe, Cu, and Mn, are involved in many redox processes that require electron transfer. Alkali and alkaline earth ions, like Na, K, and Ca, play crucial role in modulating cellular responses. Moreover metal ions like Zn, Mg, Ca, have a mainly structural role in protein and are crucial to maintaining protein folding and stabilize their structure by constraining in a proper conformation and help to fix a particular physiologically active conformation of the protein (Invernizzi et al. 2009). Among all naturally found metalloproteins, those containing Mg, Ca, and Zn appear to be the most abundant in nature. In particular among essential metals, zinc has a dominant importance (Dudev and Lim 2008) since it is an essential cofactor of many metabolic enzymes and transcription factors in proteins both for catalysis and structural functions (e.g. in superoxide dismutase, p53 and zinc-finger motifs). Indeed, zinc is one of the most abundant divalent metals in living organisms and more than 10% of human proteins and nearly half of transcription factors bind Zn^{2+} (Andreini et al. 2006), and zinc-finger domains are the most abundant structural domains in the human proteome. Zinc binding in protein can have two main roles: plays predominantly a catalytic role or serve only a structural role. The most common zinc first binding sphere found in the first category is His-3Water molecules, although catalytic sites containing His, Asp/Glu, or Cys side chains in different composition, have also been observed (Christianson and Cox 1999). The biggest class of Zn-proteins in which zinc have only structural role are those of the Zn-finger-like family, which is involved in nucleic acid binding and gene regulation, including the oncosuppressor p53 (Berg and Godwin 1997). Zn-finger-like family can be divided in classes on the basis of the metal-coordination site: the cellular or transcription factor type, characterized by a Cys_2His_2 metalbinding site (Hanas et al. 1983) the retroviral type, possessing a Cys_3His chelation sphere (Summers et al. 1990) and the steroid receptor type, having a Cys_4 metal-binding site (Petkovich et al. 1987). These classes can be further divided into different subclasses depending on the local structure and spacing between ligating residues. In contrast to the Zn-binding sites that have a catalytic role, which are in general partially exposed to solvent, the Zn-sites with structural role are deeply buried in the protein structure and are surrounded by a network of intramolecular interactions, mainly hydrogen bonds, provided by the second-coordination sphere. (Maynard and Covell 2001). Zinc prefers “soft” ligands such as Cys and His, although it is also found coordinated to Asp and Glu side chains. (Dudev et al. 2003) and the Cys side chains are considered to be deprotonated when bound to the metal. In Zn-finger-like proteins and enzymes Zn is usually tetrahedrally coordinated but it can also adopt a penta- or esa-coordinate geometry, although in aqueous solution Zn is octahedrally bound to six water molecules (Marcus, 1988). The coordination of zinc ions is essential to maintaining folding and functional conformation like in the tumor suppressor p53 and cellular levels of zinc can be altered by the presence of non-essential metal ions, like cadmium, that causes cancer (Hartwig 2010). In fact homeostasis alterations in the level of zinc in the human body are associated with development of pathological states like cancer and neurodegenerative diseases (Malgieri et al. 2011). However, little is known about the role of metals in the molecular mechanisms of protein folding, misfolding, and in the development of diseases. The structural characterization of Zn-binding proteins, and the investigation of the mechanisms of metal-mediated conformational changes, is essential in order to understand the role of metals in these biological relevant problems. Understanding the molecular basis of metal binding affinity and selectivity in metalloproteins is of fundamental importance to understand in detail the unclear mechanisms of important catalytic reactions, signal transduction, metal-induced protein folding and aggregation, heavy-metal poisoning, and metal-based therapy. Furthermore, the basic principles unravelled for protein-metal recognition would be instructive in elucidating the molecular basis of affinity/specificity in more complicated protein-protein and protein–nucleic acid recognition processes. Although a wealth of information from experiment and theory has been accumulated on protein-metal interactions the exact mechanism and physicochemical principles governing protein-metal recognition remain elusive. In this context, recent evidences showed that with the increasing accuracy and efficiency of force fields, Molecular Dynamics (MD) simulation can correctly describe complex events and can be valuable tools to elucidating the role of conformational changes in the modulation of cellular functions (Klepeis et al. 2009, Lindorff-Larsen et al. 2012). In the present project, we used these methods, to study metal-binding protein, high complex and still largely unknown systems, that have both cognitive and applicative high relevance due to the implications on human health (Loh 2010). Atomistic explicit solvent MD simulations and enhanced sampling MD techniques have been carried out on the target proteins to describe the free energy landscape associated with conformational changes and analysed by methods inspired by graph theory (Papaleo et al. 2012). We exploited MD simulations to describe long range conformational changes in p53 DNA-binding domain induced by the interaction with DNA, identifying paths of long-range communication that are likely to play an important role in promoting cofactors recruitment and in modulating p53 signalling. Our study also highlight the limitiation of current MD simulations of describing structure and dynamics of Zn-binding proteins, as well as in the description of non-essential metals as cadmium. We here propose a new approach based on molecular mechanics (MM) and quantum chemical calculations at the density functional theory (DFT) level, that permits to develop highly optimized parameters for these metal ions that can then be used in MD simulations to accurately describe the coordination of metals in metal-binding proteins.

References

Andreini, C., L. Banci, I. Bertini and A. Rosato (2006). Counting the zinc-proteins encoded in the human genome. *J Proteome Res* 5(1): 196-201.

Berg, J. M. and H. A. Godwin (1997). Lessons from zinc-binding peptides. *Annu Rev Biophys Biomol Struct* 26: 357-371.

Bertini, I., A., Sigel, H., Sigel (2002). Handbook on Metalloproteins. Biochemistry 67(7): 836-837

Christianson, D. W. and J. D. Cox (1999). Catalysis by metal-activated hydroxide in zinc and manganese metalloenzymes. *Annu Rev Biochem* 68: 33-57.

Dudev, T. and C. Lim (2008). Metal binding affinity and selectivity in metalloproteins: insights from computational studies. *Annu Rev Biophys* 37: 97-116.

Dudev, T., Y. L. Lin, M. Dudev and C. Lim (2003). First-second shell interactions in metal binding sites in proteins: a PDB survey and DFT/CDM calculations. *J Am Chem Soc* 125(10): 3168-3180.

Hanas, J. S., D. J. Hazuda, D. F. Bogenhagen, F. Y. Wu and C. W. Wu (1983). Xenopus transcription factor A requires zinc for binding to the 5 S RNA gene. *J Biol Chem* 258(23): 14120-14125.

Hartwig, A. (2010). Mechanisms in cadmium-induced carcinogenicity: recent insights. *Biometals* 23(5): 951-960.

Invernizzi, G., E. Papaleo, R. Grandori, L. De Gioia and M. Lotti (2009). Relevance of metal ions for lipase stability: structural rearrangements induced in the Burkholderia glumae lipase by calcium depletion. *J Struct Biol* 168(3): 562-570.

Klepeis, J. L., K. Lindorff-Larsen, R. O. Dror and D. E. Shaw (2009). Long-timescale molecular dynamics simulations of protein structure and function. *Curr Opin Struct Biol* 19(2): 120-127.

Lindorff-Larsen, K., N. Trbovic, P. Maragakis, S. Piana and D. E. Shaw (2012). Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *J Am Chem Soc* 134(8): 3787-3791.

Loh, S. N. (2010). The missing zinc: p53 misfolding and cancer. *Metallomics* 2(7): 442-449.

Malgieri, G., L. Zaccaro, M. Leone, E. Bucci, S. Esposito, I. Baglivo, A. Del Gatto, L. Russo, R. Scandurra, P. V. Pedone, R. Fattorusso and C. Isernia (2011). Zinc to cadmium replacement in the A. thaliana SUPERMAN Cys(2) His(2) zinc finger induces structural rearrangements of typical DNA base determinant positions. *Biopolymers* 95(11): 801-810.

Maynard, A. T. and D. G. Covell (2001). Reactivity of zinc finger cores: analysis of protein packing and electrostatic screening. *J Am Chem Soc* 123(6): 1047-1058.

Marcus Y., (1988). Ionic radii in aqueous solutions. *Chemical Reviews* 88, 1475.

Papaleo, E., K. Lindorff-Larsen and L. De Gioia (2012). Paths of long-range communication in the E2 enzymes of family 3: a molecular dynamics investigation. *Phys Chem Chem Phys* 14(36): 12515-12525.

Petkovich, M., N. J. Brand, A. Krust and P. Chambon (1987). A human retinoic acid receptor which belongs to the family of nuclear receptors. *Nature* 330(6147): 444-450.

Summers, M. F., T. L. South, B. Kim and D. R. Hare (1990). High-resolution structure of an HIV zinc fingerlike domain via a new NMR-based distance geometry approach. *Biochemistry* 29(2): 329-340.

3.1 DNA-binding protects p53 from interactions with cofactors involved in transcription-independent functions.

Matteo Lambrughi [1,2], Luca De Gioia [1], Francesco Luigi Gervasio [3], Kresten Lindorff-Larsen [2], Ruth Nussinov [4,5], Chiara Urani [6], Maurizio Bruschi [6], Elena Papaleo [2]

[1] Department of Biotechnology and Biosciences, University of Milano-Bicocca, Piazza della Scienza 2, 20126, Milan (Italy). [2]Structural Biology and NMR Laboratory, Department of Biology, University of Copenhagen, Copenhagen, Denmark. [3] Institute of Structural and Molecular Biology and Department of Chemistry, University College London, London WC1H 0AJ, United Kingdom. [4] Cancer and Inflammation Program, Leidos Biomedical Research Inc., Frederick National laboratory, National Cancer Institute, Frederick, MD 21702, USA. [5] Department of Human Molecular Genetics and Biochemistry, Sackler School of Medicine,Tel Aviv University, Tel Aviv 69978, Israel. [6] Department of Earth and Environmental Sciences, University of Milano-Bicocca, Piazza della Scienza 1, 20126, Milan (Italy)

3.1.1 Introduction

Allosteric mechanisms involve the transmission of local structural perturbations, such as those incurred by interactions with ligands, to distal protein regions (Cui and Karplus 2008, Changeux 2012, Tsai and Nussinov 2014). It has been proposed that allostery is an intrinsic feature of any protein (Gunasekaran et al. 2004, Clarkson et al. 2006) and can be described in terms of shifts in the distribution of pre-existing states present in both the free and bound protein (Kern and Zuiderweg 2003, del Sol et al. 2009). The effects induced at the distant sites can both involve major conformational changes (Bray and Duke 2004) or more subtle localized changes in either dynamics or conformation (Tsai et al. 2008, Petit et al. 2009). The transmission of the structural effects across long distances relies on dynamic coupling between different residues (Swain and Gierasch 2006, Daily et al. 2008, Goodey and Benkovic 2008, Tsai and Nussinov 2014). In this view, the network of residue contacts and the intrinsic dynamics of the protein are crucial components of long-range communication.

Given the key role that conformational changes play in the regulation of cellular events (Nussinov et al. 2013) or deregulation in disease (Peracchi and Mozzarelli 2011, Nussinov and Tsai 2013), understanding long-range structural communication in atomistic detail is a significant goal (Fenwick et al. 2011, Tzeng and Kalodimos 2011, Manley and Loria 2012). Atomistic molecular dynamics simulations (MD) have recently been successfully employed to understand mechanisms related to long-range structural communication (Fenwick et al. 2011, Collier and Ortiz 2013, Feher et al. 2014). MD can provide information on the properties of protein ensembles from femto- to milli-seconds (Dror et al. 2012) and long-range communication paths can be inferred from MD ensembles (Ghosh and Vishveshwara 2007, Collier and Ortiz 2013, Feher et al. 2014). Enhanced sampling techniques, such as metadynamics (Sutto et al. 2012), may substantially help to determine the free energy landscape associated with conformational transitions (D'Abramo et al. 2012, Palazzesi et al. 2013, Sahun-Roncero et al. 2013, Sutto and Gervasio 2013). One central hub protein for the cell that coordinates and controls several processes is the p53 tumor suppressor. The tumor suppressor p53 protein is of primary importance for human health, since it exerts an essential role in the response against DNA damage, telomere degradation and other oncogenic stress signals. p53 is a transcription factor that plays a key function in multiple anti-cancer mechanisms, through the regulation of more than 4000 genes, coordinating the pathways of control and repair of DNA, arrest and regulation of cell cycle, replicative senescence and apoptosis. The p53 acts as a molecular hub protein in cell integrating and coordinating a multitude of signaling pathways, through the recruitment of multiple interactors, in order to initiate the appropriate cellular response. This key role is demonstrated by the fact that p53 gene is mutated in more than 50% of the human tumors, making p53 as the single most frequently altered protein in cancer (Bray and Duke 2004). P53 interacts with DNA as a dimer of dimers, with its central core DNA binding domain (DBD), characterized by a β-sandwich fold, that present the coordination site for a single zinc atoms (Tsai et al. 2008). The presence of zinc is crucial for site specific DNA binding, proper transcriptional activation and tumor suppressor functions of p53 (Petit et al. 2009). Several experimental investigations reported that loss of Zn compromises DNA-binding activity, inducing misfolding and loss of stability (Clarkson et al. 2006). Despite that, lot is still unknown about the structural mechanisms that are at the basis of p53 recruitment of interactors and how these affect and regulate p53 biological functions. Indeed, p53 interacts with multiple partners and integrates diverse signalling pathways to initiate distinct cellular responses (Vogelstein et al. 2000, Vousden and Lu 2002, Joerger and Fersht 2008). The ability to exploit the same region for interaction with distinct biological partners is crucial in hub proteins and often associated with allosteric modulation (Oldfield et al. 2008, Tsai et al. 2009, Van Roey et al. 2012). Several experimental and computational studies account for p53-mediated interactions (Vogelstein et al. 2000, Joerger and Fersht 2008, Tuncbag et al. 2009, Mavinahalli et al. 2010). Nevertheless, the atom-level picture of the interactions between p53 and its biological partners is far from being complete. Indeed, we still do not know all the structures of the complexes between p53 and its biological partners, as well as we have very limited knowledge of the effects induced on p53 structure and function upon binding of other biomolecules.

The DNA-binding domain (DBD) of p53 is especially important for both cytosolic and nuclear functions of the protein (Joerger and Fersht 2008, Green and Kroemer 2009, Follis et al. 2014). Even though the p53 DNA-binding domain (DBD) does not appear to undergo a large conformational change upon binding with cofactors, a slow conformational exchange process in the proximity of the N-terminal disordered region has recently been identified by NMR experiments on the free domain in solution (Bista et al. 2012). Previous studies focused mostly on changes occurring at the DNA-binding interface and loop L1 (Canadillas et al. 2006, Pan and Nussinov 2010, Petty et al. 2011, Lukman et al. 2013), and less known about changes occurring at more distal sites. Moreover very little is known on the structural effects induced by p53 binding to the DNA and, in particular, on how the DNA-binding can affect the recruitment of specific cofactors. To gain a deeper insight into conformational changes that may underlie p53 regulation and function, we exploited all-atom explicit solvent MD simulations coupled to graph analysis and enhanced sampling by Parallel Tempering Metadynamics (PT-MetaD) (Bussi et al. 2006) performed in collaboration with Dr. Elena Papaleo at the University of Copenhagen to study p53 DBD (Figure 1A) in its DNA-bound ($p53_{DBD-DNA}$) and unbound ($p53_{DBD}$) forms. Using analysis methods inspired by graph theory, we identified a loop (loop S6-S7, residues 207-213, Figure 1A), located more than 30 Å away from the DNA-binding site, which is modulated by DNA-interaction. In particular, conformational changes in loop L1 at the DNA-binding interface are long-range coupled to changes in loop S6-S7, which is in proximity to the N-terminal disordered tail, linking our results to the slow exchange observed by NMR in the N-terminal region of the free protein in solution (Bista et al. 2012). We show that DNA binding promotes a long-range conformational change on loop S6-S7 and the coupling between the loops L1 and S6-S7 also provides a mechanism for how DNA binding may exert long-range effects and link our structural findings to p53 function. We observed a population shift toward p53DBD states that disfavour the interaction with partners involved in p53 transcription-independent apoptotic functions. In particular, DNA binding increases the population of a minor state, which is also present in the free protein in solution, at this distal site by more than four fold with respect to the free protein in solution. In this minor conformation, the interface of p53 that binds biological partners related to p53 transcription-independent functions is not accessible. The latter includes Ku70, involved in apoptotic signalling. Our study thus proposes a mechanism to protect p53 from interactions with partners that elicit its transcription-independent apoptotic signalling when p53 is bound to the DNA and has to carry out its transcription functions., such as Ku70.

3.1.2 Results and Discussion

DNA-interaction modulates long-range p53 DBD through coupling between DNA-binding loop L1 and loop S6-S7.

We firstly evaluated the stability of p53 DBD and its complex with DNA by analyzing in details the inter and intramolecular interactions (Figure 1S). The results are in great agreement with experimental data pointing out that the p53DBD and its complex with DNA are stable during our simulations.

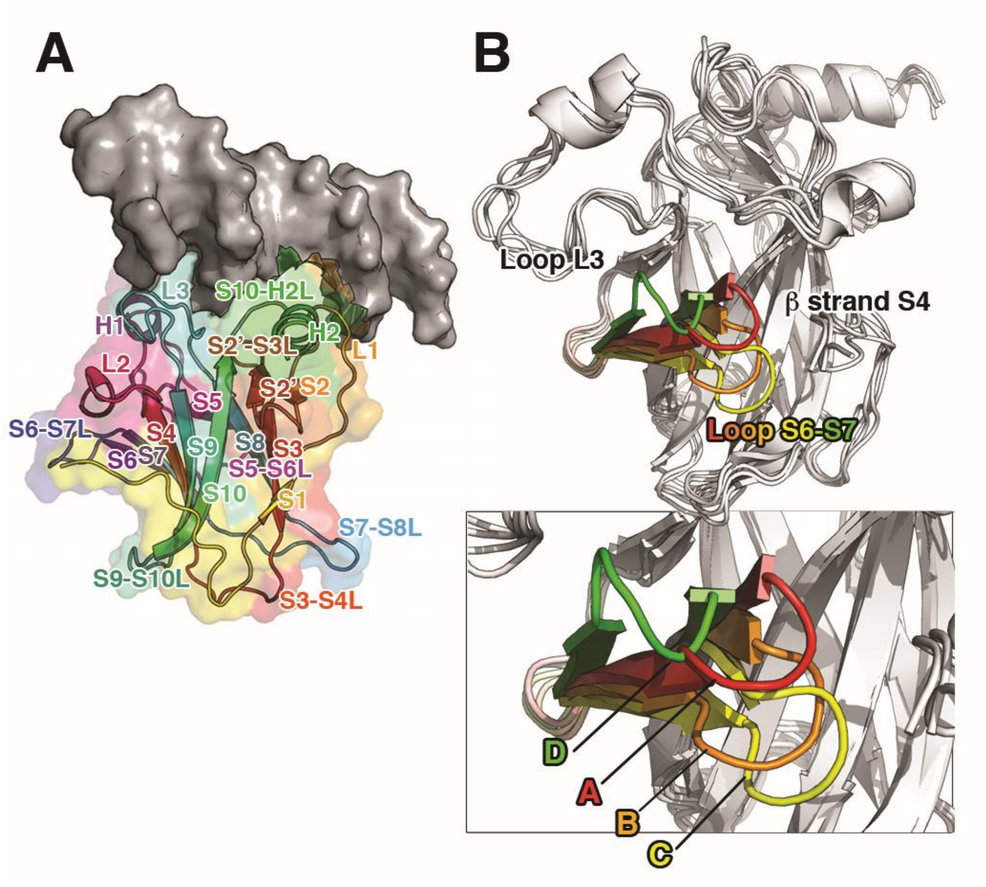


Figure 1. A) Structural features of p53 DBD. The 3D structure of p53 DBD in complex with DNA (PDB entry 1TSR) is shown as cartoon of different shade of colors from the N- (yellow) to the C-terminal extremity (green). B) Substates of S6-S7 loop in the 2D subspace described by the first and second PCs. The probability density function is represented as a contour plot. The different substates are indicated by capital letters (A-D). The average 3D structure identified for each basin is represented as cartoon and the conformations of S6-S7 loop are highlighted in red, orange, yellow and green for A, B, C and D respectively.

To identify the regions of p53DBD that are influenced by the interaction with DNA, we first performed unbiased MD simulations and a principal component analysis (PCA) of $p53_{DBD}$ and $p53_{DBD\text{-}DNA}$. Comparing the subspace described by the two first principal components (Table 1S, Figure 2 and 2S), we observe (as expected) noticeable differences in the dynamics of the loop L1 that is in direct contact with DNA (Figure 2S). In the DNA-unbound state, loop L1 samples also conformations known as 'recessed states' (Figure 3-4S). These states are known to be involved in the early events of DNA recognition (Luckman et al., 2013) as experimental studies have shown that loop L1 populates two different conformations, one extended and elongated toward the DNA, and one 'recessed' (Figure 3-4S) (Petty et al., 2011). More intriguing, we also observed differences in the dynamics of loop S6-S7 between $p53_{DBD}$ and $p53_{DBD\text{-}DNA}$ (Figure 1). This loop connects β-strands S6 and S7 (Figure 1A) and is located 3 nm away from the DNA-binding site. In the PCA subspace, we identified four major conformational substates for this loop, labelled from A to D and reported in (Figure 1B). We should note that the starting structure for the simulations has the loop S6-S7 in A-like conformation and L1 in an extended conformation.

We evaluated the time evolution of each replicate in the subspace described by the two first principal components derived by PCA based on the conformations of loop S6-S7 (Figure 2). *A* and *D* states of the loop are more populated in the absence of DNA according to our MD simulations. *C* and to a less extent *B* are more populated when DNA is bound. *E*-like conformations are partially similar to *A* states and it is not clear if they can be described as really distinct states at this stage of the work, since *E*-like structures are observed only in one MD simulation and thus no further discussed. The interaction with DNA seems to promote B- and C-like structures (Figure 2), in which loop S6 S7 gets closer to the flanking β-strand S4 (Figure 1B). To evaluate if a coupling really exists between conformational changes at the DNA-binding interface and those in loop S6-S7, we selected a subset of key residues for DNA interaction (Lys120 in L1, Arg248 in L3, Arg273 in loop S10-H2 and Arg280 in H2 α-helix) (Figure 3A) and analysed their conformations in the different PCA regions reported in Figure 2. The PCA clustering was based uniquely on the conformation of loop S6-S7. Interestingly, we can observe that different conformations of S6-S7 also select different L1 structures, ranging from L1 states not competent for DNA interaction (in A and D conformations) and DNA-bound-like states (in C) (Figure 3A). In the B conformations we observed both Lys120 states. In particular, in D and A, Lys120 side chain is displaced outward, 1 nm apart with respect to the state sampled in $p53_{DBD\text{-}DNA}$ simulations (Figure 3A, 5S). These structures also resemble experimentally observed 'recessed states' (Petty et al., 2011) (Figure 3-4S). In our simulations, Lys120 is characterized by conformations competent for DNA interaction in structures from basins C and E (Figure 3).

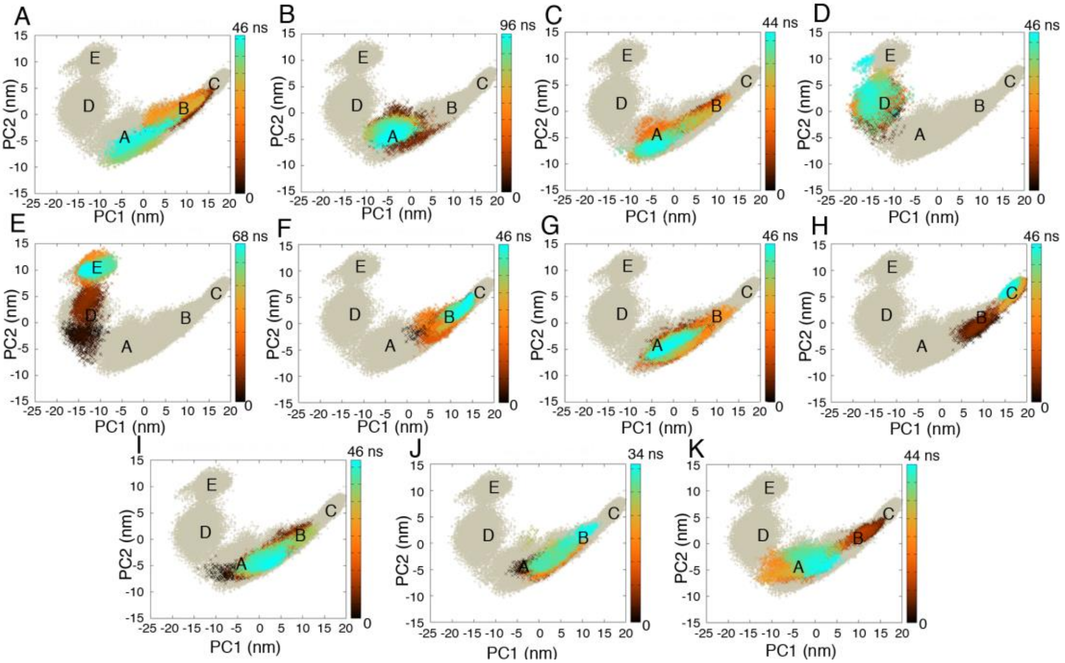


Figure 2. Time evolution of each replicate in the 2D subspace described by the first two principal components derived by PCA analysis of loop S6-S7. We mapped the space sampled by each independent replicate on the 2D projections from PCA showed in Figure 1S. The time evolution of each replicate is indicated by a color gradient, from black (0 ns) to cyan (last frame): $p53_{DBD}$ r.1 (A), r.2 (B), r.3 (C) r.4 (D) and $p53_{DBD\text{-}DNA}$ r.1 (E), r.2 (F), r.3 (G), r.4 (H), r.5 (I), r.6 (J), r.7 (K). We identified five main regions in the 2D subspace described by PCA, here labeled *A*, *B*, *C*, *D* and *E* from the more to the less populated regions.

In contrast, Arg residues at positions 248, 273 and 280 are less perturbed by DNA interaction and decoupled from S6-S7 conformations (Figure 5S), in agreement with previous computational data (Pan and Nussinov, 2010).

Both X-ray crystal structures (Natan et al. 2011) and NMR relaxation dispersion experiments (Bista et al., 2012) of a longer $p53_{DBD(91\text{-}289)}$ construct (PDB entry 2XWR) are suggestive of a role for the N-terminal tail in modulating the conformational ensemble of the p53 DBD (Figure 2B). In the $p53_{DBD(91\text{-}289)}$ structure, the N-terminal tail folds back onto the DBD interacting with Arg174 and its surroundings (Figure 3B) but it is also a region with high B-factor values. NMR relaxation dispersion experiments pointed out a slow conformational exchange involving the N-terminal tail and residues in the p53 DBD domain (Bista et al. 2012). We thus performed an additional set of MD simulations using $p53_{DBD(91\text{-}289)}$ and $p53_{DBD\text{-}DNA(91\text{-}289)}$ to test if the inclusion of the N-terminal residues can influence the dynamics of loop S6-S7. In agreement with previous data (Natan et al., 2011, Bista et al., 2012, Chillemi et al., 2013), our simulations show that the N-terminal residues can interact with DBD through a network of hydrogen bonds, including residues in the proximity of loop S6-S7 (as Thr211 and Arg213) (Figure 3B). We do not expect, however, unbiased MD simulations lasting a few hundreds of ns to provide much insight, since NMR experiments pointed out that long-timescale dynamics are involved. Indeed, even 0.5 μs MD simulations of $p53_{DBD(91-289)}$ do not show remarkable differences in the DNA-bound and free forms of S6-S7 loop (Figure 6S).

Finally, since p53 exists in solution as a tetramer (Huang et al., 2009, Bista et al., 2012), it is important to verify that the conformations of loop S6-S7 observed here are still relevant in the oligomeric assembly. Thus, we compared the structures sampled by our simulations with the known experimental quaternary structures of p53 (Figure 7S). S6-S7 conformations nicely fit in the tetrameric assembly without structural clashes. The loop residues, apart from Phe212 in some monomers, are generally solvent exposed and available for interaction with biological partners.

Overall, our data suggest that loop S6-S7 populates different states whose population is likely to be influenced by the interaction with DNA.

Long-range communication exists from loop L1 to S6-S7.

To further probe if long-range communication exists between the DNA-binding loop L1 and S6-S7 in our simulations, we employed methods inspired by graph theory to detect if there are any paths of communications between L1 and S6-S7 residues, as well as the influence of DNA on them. Paths of communications between L1 and S6-S7 loops were identified in both A-like (DNA-unbound like) and C-like structures (DNA-bound like) attesting to the existence of structural communication between these two regions, which is pre-embedded in the free protein (Figure 3C-D, Table 2S). The interaction with DNA strengthens the coupling between the two areas and also promotes new paths in the same area that were not observed in the free protein. More in details, the communication between the two loops propagates from the DNA-binding site across the external face of the core β-sheet and through loop L3 and the N-terminal tail in the $p53_{DBD-DNA}$ ensemble.

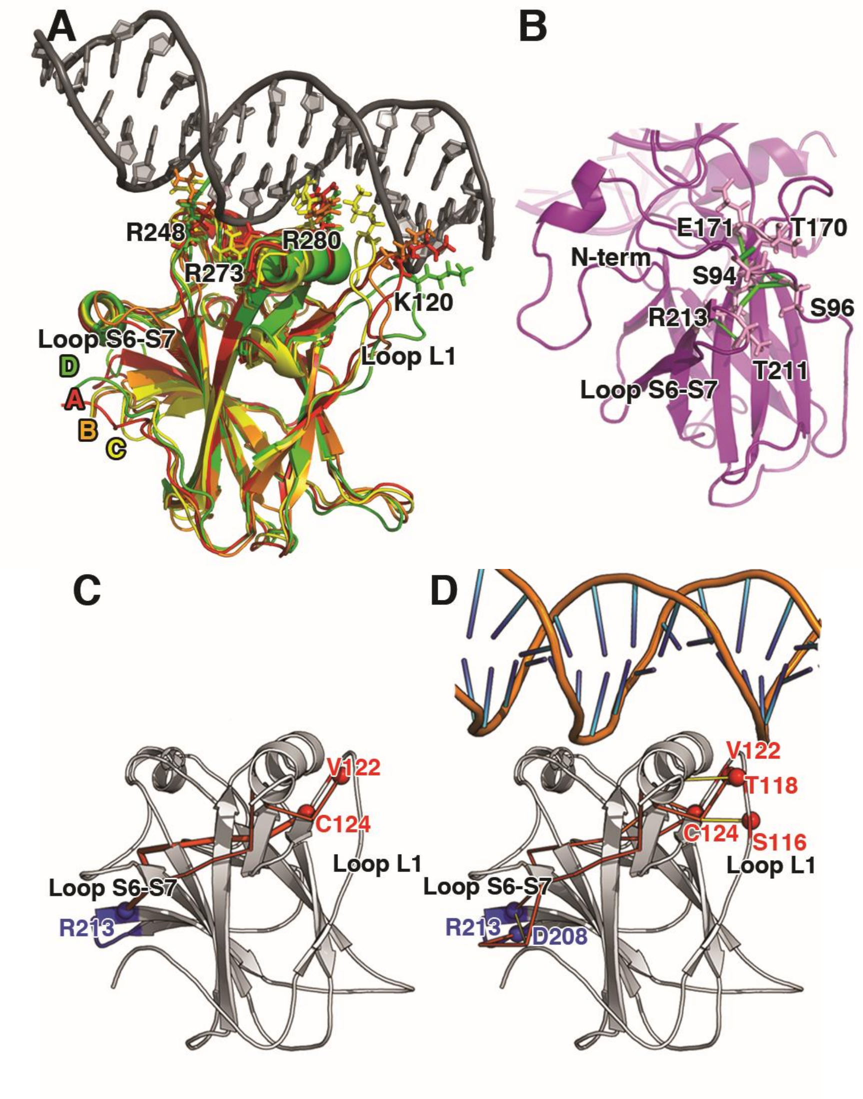


Figure 3. A) Average structure of each conformational basins of the PCA subspace. The DNA-interacting residues (Lys120, Arg248, Arg273, Arg280) are shown as sticks. We here show the coupling between the different structural states of loop S6-S7 and conformations of loop L1 ranging from states not competent for DNA interaction (as basins A and D) and DNA-bound-like states (as basins C and B). The DNA is taken from the initial structure (PDB entry 1TSR, chain B) and showed as a reference. **B) Sub-networks of hydrogen bonds in MD p53 $_{\mathbf{DBD(91\text{-}289)}}$ simulations.** We here show the cluster of hydrogen bonds that involve Thr211 and Arg213 of loop S6-S7, Ser94 and Ser90 of the N-terminal tail, along with Thr170 and Glu171 of loop L2. The Cα atoms of the residues involved in hydrogen bonds between S6-S7 loop, the N-terminal tail and other region of domain are shown as spheres. The Cα atoms of the interacting residues are connected by sticks, whose thickness is

proportional to the persistence of the interaction. The analysis was carried out by PyInteraph (Tiberti et al. 2014). **C-D) Paths of long-range communication from loop L1 to loop S6-S7.** The communication paths are shown as orange (occurrence probability > 0.25) or yellow (occurrence probability < 0.25) sticks. The terminal nodes of the paths are highlighted as blue and red spheres centred on the Cα atoms for loop S6-S7 and L1, respectively. The paths calculated from both DNA-unbound simulations (B) and DNA-bound simulations (C) are shown. Loop S6-S7 is highlighted in blue.

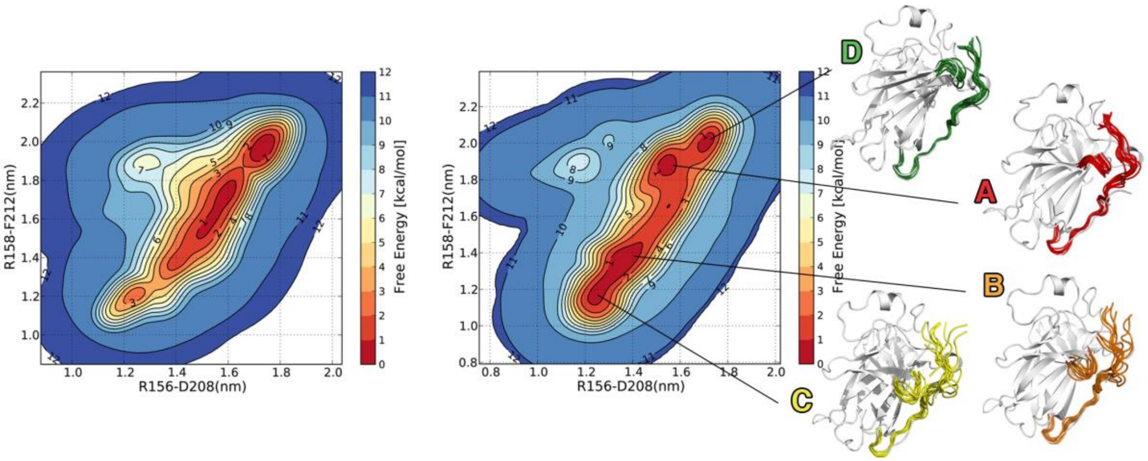


Figure 4. Interaction with DNA alters the free energy landscape of p53 DBD. The two-dimensional FES of p53DBD-DNA(91-289) is shown for two out of the four collective variables employed in PT-MetaD simulations. p53DBD(91-289) and p53DBD-DNA(91-289) are shown on the right and left panels, respectively. It is possible to observe that C-like states (the minima corresponding at lower distances in the plots) are only a minor population in the free state, whereas they are more than five fold populated when DNA is bound. The conformation of the loop and the disordered N-terminal tail in each basin of the FES are shown in red, orange, yellow and green, for A, B, C and D states respectively. The average structure for each state from the unbiased MD simulations is shown as a reference in white cartoons.

DNA-binding alters the free energy landscape of p53 DBD and promotes a population shifts in loop S6-S7

To overcome the intrinsic limitations in the conformational sampling of classical MD and obtain quantitative information on the population of the A-like/D-like states versus C-like/B-like states and the effect of the DNA binding on the free energy landscape we used enhanced-sampling techniques. We employed PT-metaD, performed in collaboration with Dr. E. Papaleo at the University of Copenhagen, with a set of four collective variables (CVs) that are based on the distances between key residues of loop S6-S7 (Arg209 and Phe212) and its surroundings (Glu258, Glu221 and Arg158). Figure 4 reports the 2D free energy surface (FES) for one pair of collective variable, whereas the other 2D FES are reported in Figure 8S. The MetaD simulations clearly show that the FES of p53 DBD is affected by DNA-interaction (Figure 4, 8S and 9S). Indeed, the interaction with DNA reshapes the free energy landscape of p53 DBD. In particular, we observed a shift in the populations of loop S6-S7 toward C-like states with a concomitant decrease in the populations of A-like and D-like states from more than 95% (free p53DBD) to less than 50-60% (DNA-bound p53DBD). When the domain is free in solution, A and D are the dominant conformations, whereas B states are mostly absent and C states are minor population of the ensemble characterized by an higher free energy with respect to the major state ($\Delta G \sim 1.7$-2 kcal/mol, Figure 9S). When DNA is bound, a decrease in the free energy barriers between the different substates of loop S6-S7 is observed and the p53 DBD is more prone to sample also B and C states, which become almost as populated as A and D (Figure 4, 8S and 9S). Loop S6-S7 is in the proximity of the N-terminal disordered region that is known to be characterized by a slow exchange for the free protein in solution (Bista et al. 2012), whereas nothing is known about the effects that DNA induced in the N-terminal tail. The exchange between these two experimentally observed states has been previously related to the breaking of the interactions between Trp91 and Arg174 (Bista et al. 2012). Indeed, the authors showed that the p53 DBD is characterized by a major state (more than 90% of the ensemble), in which the disordered tail tightly interacts with the DBD, and a minor population where these interactions are lost. Our results might suggest that changes in the hydrogen bonds mediated by S6-S7 residues could contribute to the process. According to our simulations, the loss of all these interactions not only cause the opening of the p53-hinge region located N-terminal to the DNA-binding domain (Figure 4), but it also promotes the conformational changes in loop S6-S7. We indeed show that DNA alters the free energy landscape of p53 DBD in this region and may favour the population of a minor state observed for the free ensemble in solution not only for the disordered tail, but also for the loop S6-S7.

Loop S6-S7 is a recruitment site for biological partners

To correlate the conformational changes observed in the loop S6-S7 with functional properties, we wondered if the different states of S6-S7 could act as hotspots for protein-protein interactions. We employed the Protein Interaction by Structural Matching (PRISM) method to model protein-protein interactions (Tuncbag et al. 2011). Initially, we used a subset of 40 conformations from our MD simulations of p53 DBD (95-289), both in the A-like (DNA-bound) and C-like (DNA–unbound) forms, along with the average structures from each PCA basin (Figure 1B). We included 54 different proteins in the template ensemble for PRISM (Table 3S). We selected available structures that are known to play a role in p53 pathways (Kohn, 1999) and experimentally proved to physically interact with p53 (Anderson and Appella, 2009). Our analyses pointed out several putative interaction partners (Cdk2, Crk, Chk1, RPA, Ku70, Casp3, RAP1A, Cdk7, Skp2, Cks1, Plk1, Table 3S), some of which were also previously identified (Tuncbag et al. 2009). More than 90 % of the p53-DBD conformations in the target sets were predicted to interact with 53BP2 and iASPP. These proteins are well-known interaction partners, for which experimental structures of the complexes with p53 DBD are available (PDB entries 1YCS, 2VGE) and were used here as a control. Apart from the common group of binding partners, p53 DBD A-like and C-like conformations are also predicted to interact with different biological partners as shown in Table 3S. This suggests that different conformations of loop S6-S7 could affect the spectrum of cofactors for p53 DBD and thus play a role in modulating the recruitment of different partners. In agreement with this observation, changes in loop S6-S7 from A-like to C-like conformations alter locally the solvent accessible surface, along with the electrostatic surface potential (Figure 10S). In particular, the C-like DNA-bound states feature less accessible residues in loop S6-S7 (as Asp208 and Thr211) and in the β-strand S4 (as Arg158) compared to the A-like DNA-unbound structures (Figure 10S). Among the interactors proposed by PRISM (Baspinar et al., 2014) for the p53 DBD, we selected the ones predicted to interact in the close proximity of loop S6-S7 (Table 3S). We identified protein kinases, some of which are involved in p53 regulation, especially Ark1 and Plk1. The results obtained for the kinase class of interactors are detailed in Table 3S.

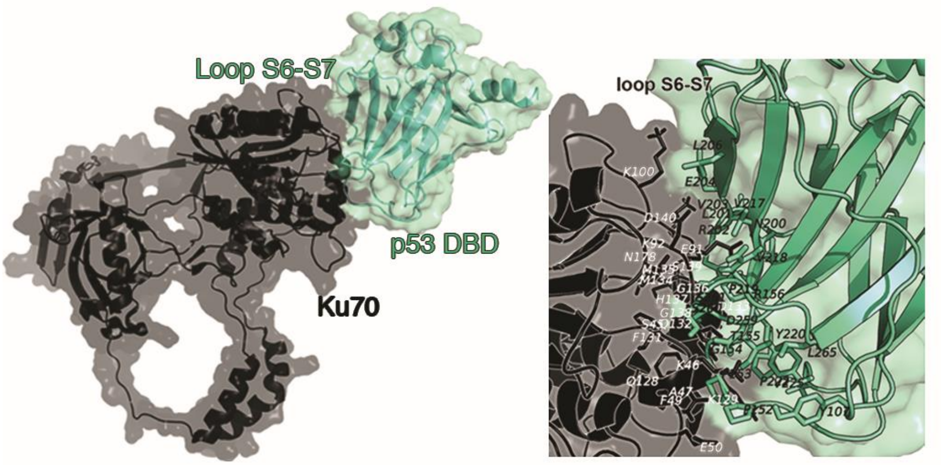


Figure 5. Complex predicted by PRIMS for Ku70 and p53 DBD. The complex predicted for interaction of Ku70 (PDB entry 1JEY, Walker et al., 2001) and p53 DBD is depicted. Structures are shown as cartoon/surface and Ku70 is highlighted in black and p53 DBD in green. The other potential interactors identified by PRISM and modulated by different S6-S7 conformations are described in details in Table 3S.

Overall, these data suggest that the conformational changes induced in the surroundings of loop S6-S7 in the DNA-bound states can protect p53 DBD from post-translational modifications. This is the case of the phosphorylation occurring at Ser215 that has been shown to inhibit p53 transcription activity (Warnock et al., 2011). Interestingly, PRISM also predicts Ku70 as a cofactor that interacts with p53 DBD only when loop S6-S7 is in the A-like DNA-unbound conformations with energy docking scores lower than -10 kcal mol-1 (Table 3S and Figure 5).

Ku70 is involved in the apoptosis control and in the p53 transcription-independent functions, mainly through formation of an inhibitory complex with Bax that is cytoprotective and impairs apoptosis initiation, preventing translocation of Bax to the mithocondria (Caelles et al., 1994). In support of the prediction, mass spectrometry and pull down assays pointed out that p53 can interact with Ku70 in vitro (Yamaguchi et al., 2009). In our predicted complex, p53 DBD interacts with the N-terminal domain of Ku70 in agreement with data showing that Bax interact with the C-terminal region of Ku70 (Sawada et al., 2003). We here show that the loop S6-S7 and its surrounding may recruit the N-terminal domain of Ku70. The interaction with DNA, according to the structures collected from our simulations, shifts the population of loop S6-S7 towards conformations not favourable for the interaction with Ku70, thus down-regulating this transcription-independent mechanism.

3.1.3 Conclusions

The p53 X-ray DBD structures available so far show very similar conformations for the DNA-bound and –unbound states (Wang et al. 2007). Nevertheless, high flexibility of p53 DBD was suggested to be crucial for its biological functions (Joerger and Fersht, 2008). Conformational changes may occur not only at the interface for DNA binding (Canadillas, 2006, Pan and Nussinov, 2010, Luckman, 2013) but also at the interface between the p53 DBD core and the disordered N-terminal region (Wells et al, 2008, Huang et al., 2009, Bista et al., 2012). Until now, MD studies have mainly focused on conformational changes in the immediate proximity of the DNA binding site (Canadillas et al., 2006, Pan and Nussinov, 2010, Petty et al., 2011, Lukman et al., 2013), neglecting changes occurring at distal sites. p53 is central in cellular signalling and interacts with many different partners to activate and regulate diverse activities and biological pathways (Vogelstein et al., 2000, Joerger and Fersht, 2008). A hub protein as p53 does not necessarily exploit different interfaces; it can also use the same binding region to interact with multiple partners (Oldfield et al., 2008, Tsai et al., 2009, Van Roey et al., 2012). Despite the central role that p53 plays in cellular functions, we have very limited knowledge of the effects induced on p53 structure and function upon binding of other biomolecules. In this context, we here show that loop S6-S7 in p53 DBD can populate different states in solution and undergoes conformational changes that can be described by simulations only when employing enhanced sampling techniques. In p53, we show that the population of loop S6-S7 states is affected by DNA through long-range effects. We observed a population shift of loop S6-S7 upon DNA binding. The structural communication between the two distal sites occurs through the L1 DNA-binding loop and the N-terminal disordered region. Mutations of residues in loop S6-S7 of p53 DBD and Arg213, in particular, are also cancer-related (Pan and Haines, 2000, Zhang et al., 2014). They are also known to affect p53 transcriptional regulation (Cho et al., 1994, Pan and Haines, 2000), suggesting that the DNA-regulated conformational changes that we observed in loop S6-S7 have even a broader functional (and dysfunctional) relevance. Indeed, DNA promotes states in which residues of loop S6-S7 lose most of the interactions with the N-terminal disordered region and interact with residues in β-strands S4 and S9, and loop L2. In those conformations, certain residues of the loops S6-S7 and S4 are shielded from the solvent, altering the accessibility of an interface region to recruit biological cofactor such as Ku70 that we predicted to interact more strongly when p53 DBD is not bound to DNA. We thus speculate that the population shift of p53DBD towards conformations less favourable for Ku70 binding could affect the signalling mediated by p53. Ku70 interaction with p53 is necessary to release and activate Bax, playing a role in the multiple steps involved in initiation of the apoptotic pathways (Speidel, 2010). The important role of the p53 DBD in apoptotic regulatory functions is also emphasized by the fact that BCL-xL interacts on the same surface of the DBD devoted to DNA-binding in the cytoplasm (Follis et al. 2014). Our study points out a new mechanism based on long-range structural communications. The conformational events in loop S6-S7 induced after DNA binding contribute, together with other effects (like post-translational modifications, to quench the interaction with Ku70 disfavoring signal transmission towards the apoptotic pathways in which Ku70 and p53 are involved.

3.1.4 Methods

Starting structures for simulations and system setup

The X-ray structure of p53 DBD (PDB entry 1TSR chain B, residues 95-289, in complex with DNA (PDB entry 1TSR chain E and F, (Cho et al. 1994) was employed as initial structure for p53DBD-DNA simulations. The same structure was used, upon DNA removal, as starting structure for p53DBD simulations. Additional simulations (p53DBD-DNA(91-289) and p53DBD(91-289)) using both MD and PT-MetaD were carried out starting from the X-ray structure of p53 DBD with four additional N-terminal residues tail (PDB entry 2XWR). The atomic coordinates of the DNA were modeled from 1TSR upon structural alignment of the two p53 DBDs. MD simulations were performed with Gromacs 4.5 (Hess et al. 2008) coupled to the PLUMED 1.3 plugin for MetaD (Bonomi et al.,2009). The systems were solvated in a dodecahedral box (minimum distance between protein and box edges: 1 nm) of Tip3p water molecules (Jorgensen et al., 1983) at 150 mM NaCl, applying periodic boundary conditions. We used the CHARMM22/CMAP force field (Mackerell et al. 2004). We collected overall 25 independent MD simulations (replicates) of 50-100 ns each (Table 1S) at 300 K and 1 atm using the isothermal-isobaric ensemble (NPT) and an external Berendsen bath with thermal and pressure coupling of 0.1 and 1 ps, respectively. The non-bonded interaction list was updated every 10 steps and conformations stored every 4 ps in the productive MD runs.

Classical MD simulations.

Each system was initially relaxed by 10000 steps of energy minimization by the steepest descent method. The optimization step was followed by 50 ps of solvent equilibration at 300K, while restraining the protein atomic positions using a harmonic potential. Each system was then equilibrated to the target temperature (300 K) and pressure (1 bar) through thermalization and a series of pressurization simulations of 100 ps each. We performed productive MD simulations using LINCS algorithm (Hess et al. 1993) to constrain heavy-atom bonds, allowing for a 2 fs time-step. Long-range electrostatic interactions were described by the Particle-mesh Ewald summation scheme (Darden et al. 1993). Van der Waals and Coulomb interactions were truncated at 0.9 nm. We evaluated the radius of gyration and the secondary structure content to ensure that they do not dramatically deviates from the corresponding values in the known experimental structures (PDB entries: 1TSR, 2XWR, 4HJE). We also evaluated the evolution of the main-chain root mean square deviation (rmsd) with respect to the initial structure over the simulation time (Figure 11S). The main-chain rmsd required up to five ns, which were thus discarded from further analyses, to reach stable values. We also evaluated the distances between the Zn^{2+} metal ion and its coordinating residues (Cys176, His179, Cys238 and Cys242) over the simulation time (Figure 12S). To achieve an overall description of the conformational space described by our MD simulations, we merged all the structures sampled by independent replicates of the same system in a concatenated trajectory ($p53_{DBD}$ and $p53_{DBD-DNA}$ for free p53 DBD and DBD in complex with DNA, respectively).

Enhanced sampling

We used PT-MetaD (Bussi et al. 2006) in the well-tempered ensemble (WTE) (Bonomi et al. 2009) to enhance sampling. In such simulations, sampling of the free energy surface is enhanced by adding a history-dependent potential to a set of collective variables (CVs). We also further enhanced the sampling by allowing for exchanges between different temperatures through a replica exchange approach. We employed seven replicas (at 296K, 300K, 308K, 320K, 332K, 345K, 358K) where the width of the energy distribution (of all but the “neutral” 298K replica) was increased as previously described (Sutto et al., 2013). All the replicas are subject to an additional biasing force through WTE-MetaD simulations in which a Gaussian of width 0.1 nm in all the four dimensions is deposited in the collective variable space every 4 ps (initial height of 0.5 kJ mol-1 and bias factor of 6). The set of CVs are Cα-Cα distance that are described in Text 1S. PT-MetaD/WTE simulations were carried out for 0.3 μs per replica (aggregate time 2.1 μs). Convergence of the simulations is discussed in Figure 8S.

PT-MetaD.

We used the WTE approach to overcome the inherently problematic scaling properties of PT simulations with a large number of atoms, which typically require many close temperatures to ensure sufficient energy overlap between adjacent replicas hence increasing the exchange rate. In the WTE approach, a constant bias on the energy is added to each replica to increase the width of the energy distribution so that a suitable exchange rate is ensured even when a more widely spaced temperature scheme is used. In our PT-metaD/WTE simulations, the van der Waals interactions were smoothly shifted to zero between 0.8 and 1.0 nm, and the long-range electrostatic interactions were calculated by the PME algorithm with a 0.12 nm mesh spacing combined with a switch function for the direct space between 0.8 and 1.0 nm. The system evolved in the canonical ensemble, coupled with a velocity-rescale thermostat (Bussi et al. 2007) and a time step of 2 fs. We employed, as collective variables (CVs), four pairwise Cα-Cα distances between residues of the loop S6-S7 and its surroundings: Asp208-Arg156 (CV 1), Arg158-Phe212 (CV 2), Arg209-Glu221 (CV 3) and Arg209-Glu258 (CV 4). We have selected this set of CVs as the most representative of the loop substates upon analyses of the distances distributions between all the Cα-Cα contacts that the loop makes with its neighbors in the unbiased classical MD simulations.

Paths of communication

The shortest paths of communication between Lys120 and residues of loop S6-S7 were calculated from the unbiased MD simulations using the Protein Structure Network/Linear Mutual Information (PSN/LMI) method (Lange and Grubmuller 2006, Angelova et al. 2011). Briefly, the Cα LMI matrix was calculated by averaging over non-overlapping windows of 1 ns, and using a cut-off of 0.4 to reduce noise. The Icrit value employed in our simulations was 7, calculated as previously described (Ghosh and Vishveshwara 2007, Papaleo et al. 2012, Invernizzi et al., 2014). The PSN was calculated for each structure and only edges present in at least half of the simulation frames were considered. The PSN, LMI and PSN-LMI calculations were performed using WORDOM (Seeber et al. 2011). The plots of the paths on the 3D structures were obtained by the xPyder (Pasi et al. 2012) plugin for PyMOL.

Protein Interactions by Structural Matching (PRISM)

PRISM is a motif-based protein interaction prediction software (Tuncbag et al. 2011). It includes a rigid-body structural comparison of target proteins to known templates of protein-protein interfaces and a flexible refinement using a docking energy function. We employed the standalone version of PRISM (Tuncbag et al. 2011) to screen 54 different potential interactors for the p53 DBD. We used 40 conformations of p53 DBD from our simulations, including both the DNA-bound and unbound states, along with the average structures from each PCA basin. The potential interactors selected at the first round step were then submitted to the PRISM2.0 web server (Baspinar et al. 2014) and evaluated on the basis of docking energy scores.

3.1.5 Acknowledgements

E. P. and K. L.-L. were supported by a Hallas-Møller stipend from the Novo Nordisk Foundation. The results of this research have been achieved using the PRACE-DECI project DyNet granted to E.P and the ISCRA-CINECA HPC Grants (HP10BLFPW4 and HP10C8LO8N) to E.P and L.D.G. The authors also thank Pierfrancesco Pagella and Elisabetta Morandi for their collaboration in the first stages of the project and Gaetano Invernizzi for careful reading of the manuscript and fruitful discussion. This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract number HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. This research was supported (in part) by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

3.1.6 References

Anderson, C., E., Appella (2009) Signaling to the p53 tumor suppressor through pathways activated by genotoxic and non-genotoxic stresses. *Handb Cell Signal* 264:2185–2203.

Angelova, K., A. Felline, M. Lee, M. Patel, D. Puett and F. Fanelli (2011). Conserved amino acids participate in the structure networks deputed to intramolecular communication in the lutropin receptor. *Cell Mol Life Sci* 68(7): 1227-1239.

Baspinar, A., E. Cukuroglu, R. Nussinov, O. Keskin and A. Gursoy (2014). PRISM: a web server and repository for prediction of protein-protein interactions and modeling their 3D complexes. *Nucleic Acids Res* 42(Web Server issue): W285-289.

Bista, M., S. M. Freund and A. R. Fersht (2012). Domain-domain interactions in full-length p53 and a specific DNA complex probed by methyl NMR spectroscopy. *Proc Natl Acad Sci U S A* 109(39): 15752-15756.

Bonomi, M., A. Barducci and M. Parrinello (2009). Reconstructing the equilibrium Boltzmann distribution from well-tempered metadynamics. *J Comput Chem* 30(11): 1615-1621.

Bonomi M et al. (2009) PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Comput Phys Commun* 180:1961–1972.

Bray, D. and T. Duke (2004). Conformational spread: the propagation of allosteric states in large multiprotein complexes. *Annu Rev Biophys Biomol Struct* 33: 53-73.

Bussi, G., F. L. Gervasio, A. Laio and M. Parrinello (2006). Free-energy landscape for beta hairpin folding from combined parallel tempering and metadynamics. *J Am Chem Soc* 128(41): 13435-13441.

Bussi, G., D., Donadio, M., Parrinello (2007) Canonical sampling through velocity rescaling. *J ChemPhys* 126:014101.

Caelles, C., A. Helmberg and M. Karin (1994). p53-dependent apoptosis in the absence of transcriptional activation of p53-target genes. *Nature* 370(6486): 220-223.

Canadillas, J. M., H. Tidow, S. M. Freund, T. J. Rutherford, H. C. Ang and A. R. Fersht (2006). Solution structure of p53 core domain: structural basis for its instability. *Proc Natl Acad Sci U S A* 103(7): 2109-2114.

Changeux, J. P. (2012). Allostery and the Monod-Wyman-Changeux model after 50 years. *Annu Rev Biophys* 41: 103-133.

Chillemi, G., P. Davidovich, M. D'Abramo, T. Mametnabiev, A. V. Garabadzhiu, A. Desideri and G. Melino (2013). Molecular dynamics of the full-length p53 monomer. *Cell Cycle* 12(18): 3098-3108.

Cho, Y., S. Gorina, P. D. Jeffrey and N. P. Pavletich (1994). Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. *Science* 265(5170): 346-355.

Clarkson, M. W., S. A. Gilmore, M. H. Edgell and A. L. Lee (2006). Dynamic coupling and allosteric behavior in a nonallosteric protein. *Biochemistry* 45(25): 7693-7699.

Collier, G. and V. Ortiz (2013). Emerging computational approaches for the study of protein allostery. *Arch Biochem Biophys* 538(1): 6-15.

Cui, Q. and M. Karplus (2008). Allostery and cooperativity revisited. *Protein Sci* 17(8): 1295-1307.

D'Abramo, M., O. Rabal, J. Oyarzabal and F. L. Gervasio (2012). Conformational selection versus induced fit in kinases: the case of PI3K-gamma. *Angew Chem Int Ed Engl* 51(3): 642-646.

Daily, M. D., T. J. Upadhyaya and J. J. Gray (2008). Contact rearrangements form coupled networks from local motions in allosteric proteins. *Proteins* 71(1): 455-466.

Darden, T., D., York, L., Pedersen (1993) Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J Chem Phys* 98:10089.

Del Sol, A., C. J. Tsai, B. Ma and R. Nussinov (2009). The origin of allosteric functional modulation: multiple pre-existing pathways. *Structure* 17(8): 1042-1050.

Dror, R. O., R. M. Dirks, J. P. Grossman, H. Xu and D. E. Shaw (2012). Biomolecular simulation: a computational microscope for molecular biology. *Annu Rev Biophys* 41: 429-452.

Feher, V. A., J. D. Durrant, A. T. Van Wart and R. E. Amaro (2014). Computational approaches to mapping allosteric pathways. *Curr Opin Struct Biol* 25: 98-103.

Fenwick, R. B., S. Esteban-Martin and X. Salvatella (2011). Understanding biomolecular motion, recognition, and allostery by use of conformational ensembles. *Eur Biophys J* 40(12): 1339-1355.

Follis, A. V., F. Llambi, L. Ou, K. Baran, D. R. Green and R. W. Kriwacki (2014). The DNA-binding domain mediates both nuclear and cytosolic functions of p53. *Nat Struct Mol Biol* 21(6): 535-543.

Ghosh, A. and S. Vishveshwara (2007). A study of communication pathways in methionyl- tRNA synthetase by molecular dynamics simulations and structure network analysis. *Proc Natl Acad Sci U S A* 104(40): 15711-15716.

Goodey, N. M. and S. J. Benkovic (2008). Allosteric regulation and catalysis emerge via a common route. *Nat Chem Biol* 4(8): 474-482.

Green, D. R. and G. Kroemer (2009). Cytoplasmic functions of the tumour suppressor p53. *Nature* 458(7242): 1127-1130.

Gunasekaran, K., B. Ma and R. Nussinov (2004). Is allostery an intrinsic property of all dynamic proteins? *Proteins* 57(3): 433-443.

Hess, B., H., Bekker, H., Berendsen, J., Fraaije (1993) LINCS: A linear constraint solver for molecular simulations., *J Comput Chem*, 12: 1463-1472.

Hess, B., C., Kutzner, D., van der Spoel, E., Lindahl (2008) GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J Chem Theory Comput* 4:435–447.

Huang, F., S. Rajagopalan, G. Settanni, R. J. Marsh, D. A. Armoogum, N. Nicolaou, A. J. Bain, E. Lerner, E. Haas, L. Ying and A. R. Fersht (2009). Multiple conformations of full-length p53 detected with single-molecule fluorescence resonance energy transfer. *Proc Natl Acad Sci U S A* 106(49): 20758-20763.

Invernizzi, G., M. Tiberti, M. Lambrughi, K. Lindorff-Larsen and E. Papaleo (2014). Communication routes in ARID domains between distal residues in helix 5 and the DNA-binding loops. *PLoS Comput Biol* 10(9): e1003744.

Joerger, A. C. and A. R. Fersht (2008). Structural biology of the tumor suppressor p53. *Annu Rev Biochem* 77: 557-582.

Jorgensen, W.L., J., Chandrasekhar, J. D., Madura, R. W., Impey, M. L., Klein (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79:926.

Kern, D. and E. R. Zuiderweg (2003). The role of dynamics in allosteric regulation. *Curr Opin Struct Biol* 13(6): 748-757.

Kohn, K. W. (1999). Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol Biol Cell* 10(8): 2703-2734.

Lange, O. F. and H. Grubmuller (2006). Generalized correlation for biomolecular dynamics. *Proteins* 62(4): 1053-1061.

Lukman, S., D. P. Lane and C. S. Verma (2013). Mapping the structural and dynamical features of multiple p53 DNA binding domains: insights into loop 1 intrinsic dynamics. *PLoS One* 8(11): e80221.

Mackerell, A. D., Jr., M. Feig and C. L. Brooks, 3rd (2004). Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem* 25(11): 1400-1415.

Manley, G. and J. P. Loria (2012). NMR insights into protein allostery. *Arch Biochem Biophys* 519(2): 223-231.

Mavinahalli, J. N., A. Madhumalar, R. W. Beuerman, D. P. Lane and C. Verma (2010). Differences in the transactivation domains of p53 family members: a computational study. *BMC Genomics* 11 Suppl 1: S5.

Natan, E., C. Baloglu, K. Pagel, S. M. Freund, N. Morgner, C. V. Robinson, A. R. Fersht and A. C. Joerger (2011). Interaction of the p53 DNA-binding domain with its n-terminal extension modulates the stability of the p53 tetramer. *J Mol Biol* 409(3): 358-368.

Nussinov, R. and C. J. Tsai (2013). Allostery in disease and in drug discovery. *Cell* 153(2): 293-305.

Nussinov, R., C. J. Tsai and B. Ma (2013). The underappreciated role of allostery in the cellular network. *Annu Rev Biophys* 42: 169-189.

Oldfield, C. J., J. Meng, J. Y. Yang, M. Q. Yang, V. N. Uversky and A. K. Dunker (2008). Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics* 9 Suppl 1: S1.

Palazzesi, F., A. Barducci, M. Tollinger and M. Parrinello (2013). The allosteric communication pathways in KIX domain of CBP. *Proc Natl Acad Sci U S A* 110(35): 14237-14242.

Pan, Y. and R. Nussinov (2010). Lysine120 interactions with p53 response elements can allosterically direct p53 organization. *PLoS Comput Biol* 6(8).

Papaleo, E., K. Lindorff-Larsen and L. De Gioia (2012). Paths of long-range communication in the E2 enzymes of family 3: a molecular dynamics investigation. *Phys Chem Chem Phys* 14(36): 12515-12525.

Papaleo, E., L., Sutto, F. L., Gervasio, K., Lindorff-Larsen (2014) Conformational Changes and Free Energies in a Proline Isomerase. *J Chem Theory Comput*10:4169.

Pasi, M., M. Tiberti, A. Arrigoni and E. Papaleo (2012). xPyder: a PyMOL plugin to analyze coupled residues and their networks in protein structures. *J Chem Inf Model* 52(7): 1865-1874.

Peracchi, A. and A. Mozzarelli (2011). Exploring and exploiting allostery: Models, evolution, and drug targeting. *Biochim Biophys Acta* 1814(8): 922-933.

Petit, C. M., J. Zhang, P. J. Sapienza, E. J. Fuentes and A. L. Lee (2009). Hidden dynamic allostery in a PDZ domain. *Proc Natl Acad Sci U S A* 106(43): 18249-18254.

Petty, T. J., S. Emamzadah, L. Costantino, I. Petkova, E. S. Stavridi, J. G. Saven, E. Vauthey and T. D. Halazonetis (2011). An induced fit mechanism regulates p53 DNA binding kinetics to confer sequence specificity. *EMBO J* 30(11): 2167-2176.

Sahun-Roncero, M., B. Rubio-Ruiz, G. Saladino, A. Conejo-Garcia, A. Espinosa, A. Velazquez-Campoy, F. L. Gervasio, A. Entrena and R. Hurtado-Guerrero (2013). The mechanism of allosteric coupling in choline kinase alpha1 revealed by the action of a rationally designed inhibitor. *Angew Chem Int Ed Engl* 52(17): 4582-4586.

Sawada, M., W. Sun, P. Hayes, K. Leskov, D. A. Boothman and S. Matsuyama (2003). Ku70 suppresses the apoptotic translocation of Bax to mitochondria. *Nat Cell Biol* 5(4): 320-329.

Seeber, M., A. Felline, F. Raimondi, S. Muff, R. Friedman, F. Rao, A. Caflisch and F. Fanelli (2011). Wordom: a user-friendly program for the analysis of molecular structures, trajectories, and free energy surfaces. *J Comput Chem* 32(6): 1183-1194.

Speidel, D. (2010). Transcription-independent p53 apoptosis: an alternative route to death. *Trends Cell Biol* 20(1): 14-24.

Sutto, L., M. Suppo, F.L. Gervasio (2012). New advances in metadynamics. *Wiley Interdiscip Rev Comput Mol Sci*(2): 771-779.

Sutto, L. and F. L. Gervasio (2013). Effects of oncogenic mutations on the conformational free-energy landscape of EGFR kinase. *Proc Natl Acad Sci U S A* 110(26): 10616-10621.

Swain, J. F. and L. M. Gierasch (2006). The changing landscape of protein allostery. *Curr Opin Struct Biol* 16(1): 102-108.

Tiberti, M., G. Invernizzi, M. Lambrughi, Y. Inbar, G. Schreiber and E. Papaleo (2014). PyInteraph: a framework for the analysis of interaction networks in structural ensembles of proteins. *J Chem Inf Model* 54(5): 1537-1551.

Tsai, C. J., A. del Sol and R. Nussinov (2008). Allostery: absence of a change in shape does not imply that allostery is not at play. *J Mol Biol* 378(1): 1-11.

Tsai, C. J., B. Ma and R. Nussinov (2009). Protein-protein interaction networks: how can a hub protein bind so many different partners? *Trends Biochem Sci* 34(12): 594-600.

Tsai, C. J. and R. Nussinov (2014). A unified view of "how allostery works". *PLoS Comput Biol* 10(2): e1003394.

Tuncbag, N., A. Gursoy, R. Nussinov and O. Keskin (2011). Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nat Protoc* 6(9): 1341-1354.

Tuncbag, N., G. Kar, A. Gursoy, O. Keskin and R. Nussinov (2009). Towards inferring time dimensionality in protein-protein interaction networks by integrating structures: the p53 example. *Mol Biosyst* 5(12): 1770-1778.

Tzeng, S. R. and C. G. Kalodimos (2011). Protein dynamics and allostery: an NMR view. *Curr Opin Struct Biol* 21(1): 62-67.

Van Roey, K., T. J. Gibson and N. E. Davey (2012). Motif switches: decision-making in cell regulation. *Curr Opin Struct Biol* 22(3): 378-385.

Vogelstein, B., D. Lane and A. J. Levine (2000). Surfing the p53 network. *Nature* 408(6810): 307-310.

Vousden, K. H. and X. Lu (2002). Live or let die: the cell's response to p53. *Nat Rev Cancer* 2(8): 594-604.

Walker, J. R., R. A. Corpina and J. Goldberg (2001). Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. *Nature* 412(6847): 607-614.

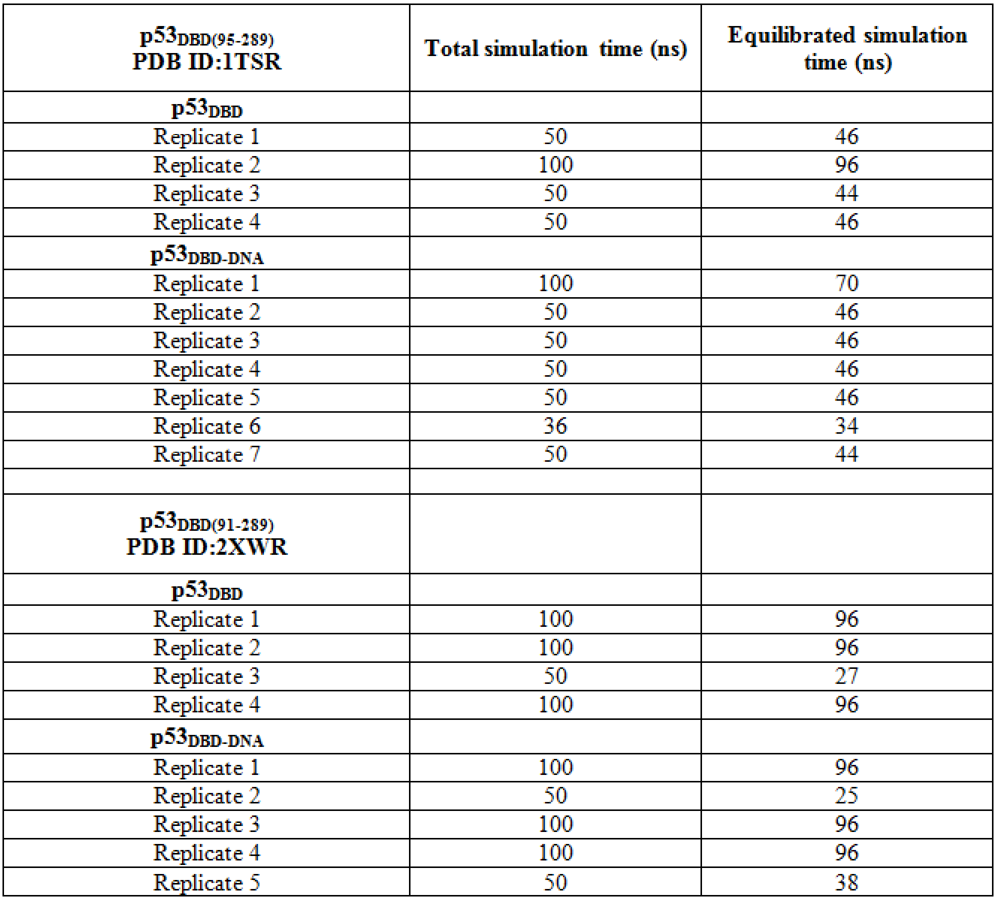
Wang, Y., A. Rosengarth and H. Luecke (2007). Structure of the human p53 core domain in the absence of DNA. *Acta Crystallogr D Biol Crystallogr* 63(Pt 3): 276-281.

Wells, M., H. Tidow, T. J. Rutherford, P. Markwick, M. R. Jensen, E. Mylonas, D. I. Svergun, M. Blackledge and A. R. Fersht (2008). Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proc Natl Acad Sci U S A* 105(15): 5762-5767.

Yamaguchi, H., N. T. Woods, L. G. Piluso, H. H. Lee, J. Chen, K. N. Bhalla, A. Monteiro, X. Liu, M. C. Hung and H. G. Wang (2009). p53 acetylation is crucial for its transcription-independent proapoptotic functions. *J Biol Chem* 284(17): 11171-11183.

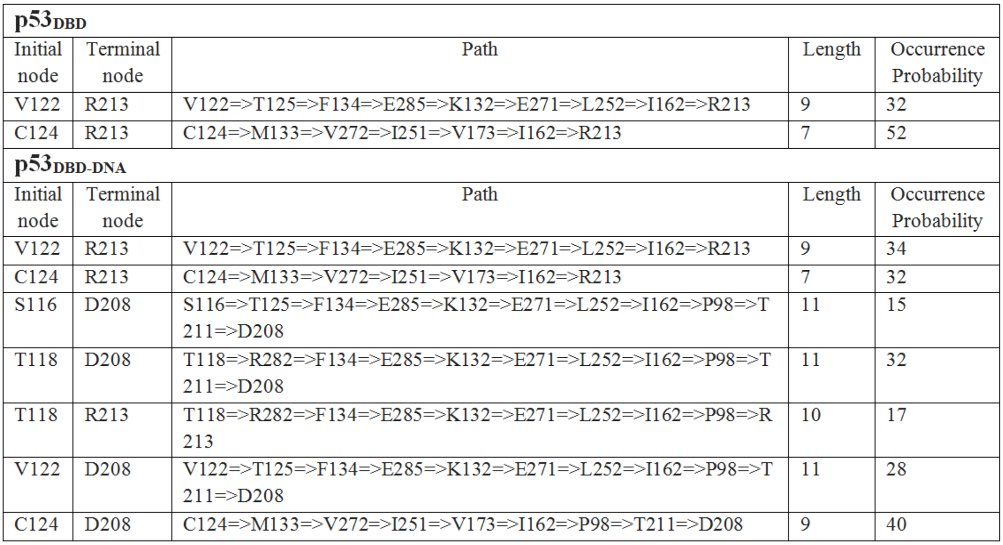
3.1.7 Supplementary Materials

Table 1S Summary of the unbiased classical MD simulations.



p53$_{\text{DBD(95-289)}}$ PDB ID:1TSR	Total simulation time (ns)	Equilibrated simulation time (ns)
p53$_{\text{DBD}}$		
Replicate 1	50	46
Replicate 2	100	96
Replicate 3	50	44
Replicate 4	50	46
p53$_{\text{DBD-DNA}}$		
Replicate 1	100	70
Replicate 2	50	46
Replicate 3	50	46
Replicate 4	50	46
Replicate 5	50	46
Replicate 6	36	34
Replicate 7	50	44
p53$_{\text{DBD(91-289)}}$ PDB ID:2XWR		
p53$_{\text{DBD}}$		
Replicate 1	100	96
Replicate 2	100	96
Replicate 3	50	27
Replicate 4	100	96
p53$_{\text{DBD-DNA}}$		
Replicate 1	100	96
Replicate 2	50	25
Replicate 3	100	96
Replicate 4	100	96
Replicate 5	50	38

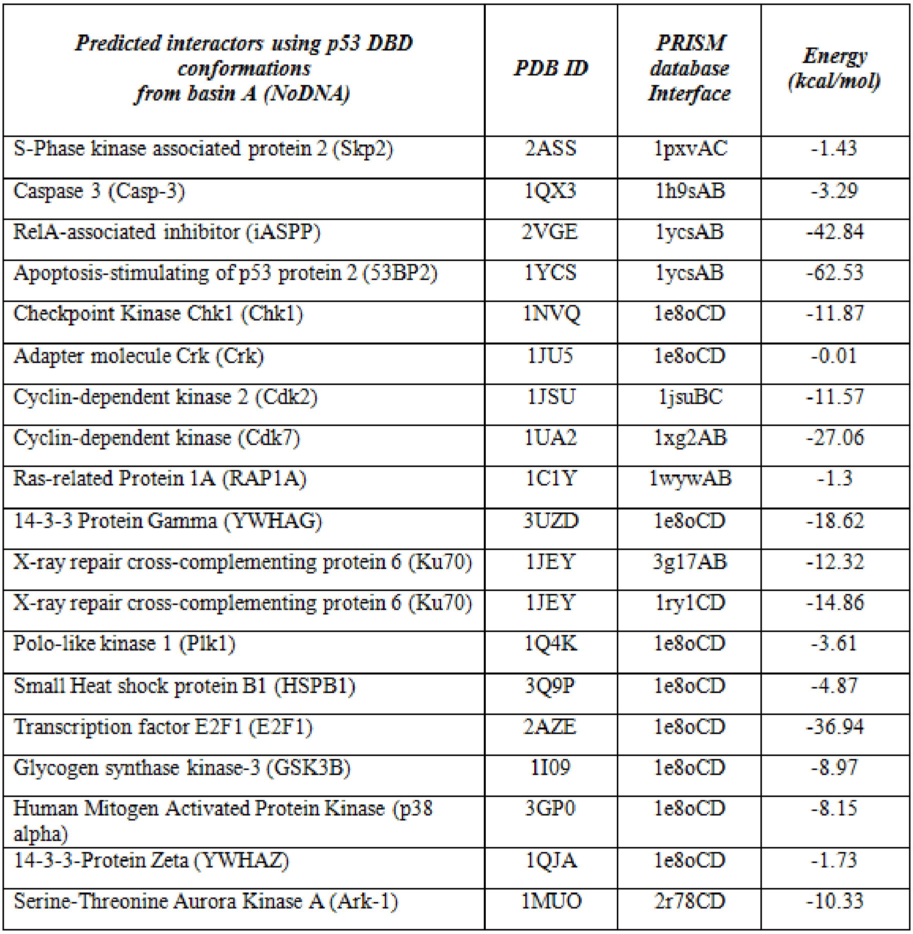
Table 2S. Paths of long-range communication in p53$_{DBD}$ p53$_{DBD-DNA}$. The paths depicted in Figure 2C-D are listed below



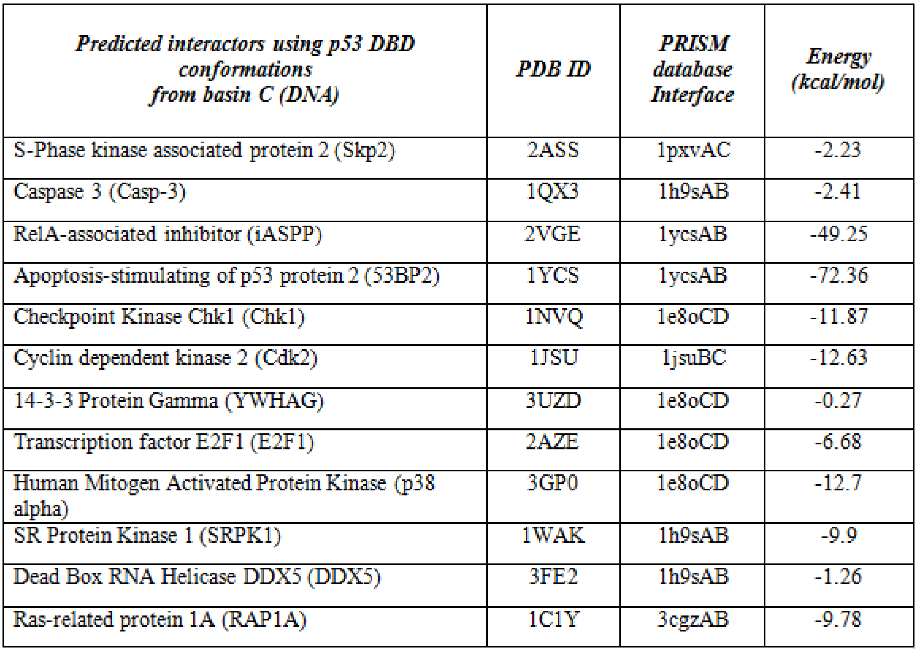
p53$_{DBD}$				
Initial node	Terminal node	Path	Length	Occurrence Probability
V122	R213	V122=>T125=>F134=>E285=>K132=>E271=>L252=>I162=>R213	9	32
C124	R213	C124=>M133=>V272=>I251=>V173=>I162=>R213	7	52
p53$_{DBD-DNA}$				
Initial node	Terminal node	Path	Length	Occurrence Probability
V122	R213	V122=>T125=>F134=>E285=>K132=>E271=>L252=>I162=>R213	9	34
C124	R213	C124=>M133=>V272=>I251=>V173=>I162=>R213	7	32
S116	D208	S116=>T125=>F134=>E285=>K132=>E271=>L252=>I162=>P98=>T211=>D208	11	15
T118	D208	T118=>R282=>F134=>E285=>K132=>E271=>L252=>I162=>P98=>T211=>D208	11	32
T118	R213	T118=>R282=>F134=>E285=>K132=>E271=>L252=>I162=>P98=>R213	10	17
V122	D208	V122=>T125=>F134=>E285=>K132=>E271=>L252=>I162=>P98=>T211=>D208	11	28
C124	D208	C124=>M133=>V272=>I251=>V173=>I162=>P98=>T211=>D208	9	40

Table 3S PRISM prediction of p53 binding partners. The interactors predicted by PRISM using p53 DBD *A* and *C* conformations are listed below. If more than one interaction interface was identified by PRISM for the same interactor-p53$_{DBD}$ complex, we took into account the one predicted with the minimum energy value. The predicted interactors are mainly involved in the regulation of p53 stability and activation, as several kinases like Chk1, Cdk2, Ark-1, Plk1, or in the apoptotic pathways, as Ku70 and Casp-3 (Bista et al., 2012). Skp2, Casp-3, iASPP, 53BP2, Chk1, Cdk2, YWHAG, E2F1, p38 alpha and Rap1A can equally bound both *A*-like structures (more populated in the free domain) and *C*-like structures (more populated in the DNA-bound state). DDX5 and SRPK1 are predicted to interact only with *C*-like conformations. Loop S6-S7 is not directly involved, however, in the interactions with DDX5 and SRPK1 in the model predicted by PRISM. They are covering the upper region with respect to it, including the loop L3, suggesting that also fluctuations in L3 region might play a role for cofactor recruitment. Ku70, HSPB1, Crk, Cdk7, Plk1, GSK3B and Ark-1 are predicted to interact preferentially with DNA-unbound *A*-like conformations in the proximity of S6-S7 loop. The protein kinases Plk-1 and Ark-1 bind the region of p53 DBD below the S6-S7 loop, comprising some β-strands (S4, S7 and S9), along with residues from loops S9-S10 and S3-S4. Several of the interactors predicted to be recruited only by p53 *A*-like states are protein kinases. In this context, we noticed that one phosphorylation site for Ark-1 is present in the p53 DBD (Ser215) close to the loop S6-S7. The phosphorylation on Ser215, which is localized in the β-sheet S7, is mediated by the serine/threonine kinases Aurora A, i.e. Ark-1, which is here identified as a p53 interactor (Canadillas eet al., 2006).

Phosphorylation of Ser215 is suggested to inactivates p53 and alters its transcription regulatory activity (Carmena and Earnshw, 2003). In the DNA-bound *C*-like states some of the residues in the surroundings of Ser215, as for example Asp208 and Arg158, are less solvent-accessible (SAS from 0.7 to 0.3 nm^2 in *A*-like and *C*-like states, respectively) possibly contributing to reduce the accessibility of the phospho-site residue. Our results thus suggest that DNA-bound states unfavour post-translational modification on Ser215 of the p53 DBD targeting for p53 inactivation, thus playing a role in its regulation. Two different complexes were predicted for the interaction between p53 DBD and Ku70. Both the poses were identified only when p53 is in *A*-like conformations (i.e. DNA-unbound) and interacting with the von Willebrand A domain of Ku70 but in slightly different regions. We selected only one of the complexes (3g17AB interface from PRISM database) since the second pose (1ry1CD interface) was not reliable due to missing coordinates for residues at the interface in the X-ray structure of Ku70 (PDB entry 1JEY, region 223-231).



Predicted interactors using p53 DBD conformations from basin A (NoDNA)	*PDB ID*	*PRISM database Interface*	*Energy (kcal/mol)*
S-Phase kinase associated protein 2 (Skp2)	2ASS	1pxvAC	-1.43
Caspase 3 (Casp-3)	1QX3	1h9sAB	-3.29
RelA-associated inhibitor (iASPP)	2VGE	1ycsAB	-42.84
Apoptosis-stimulating of p53 protein 2 (53BP2)	1YCS	1ycsAB	-62.53
Checkpoint Kinase Chk1 (Chk1)	1NVQ	1e8oCD	-11.87
Adapter molecule Crk (Crk)	1JU5	1e8oCD	-0.01
Cyclin-dependent kinase 2 (Cdk2)	1JSU	1jsuBC	-11.57
Cyclin-dependent kinase (Cdk7)	1UA2	1xg2AB	-27.06
Ras-related Protein 1A (RAP1A)	1C1Y	1wywAB	-1.3
14-3-3 Protein Gamma (YWHAG)	3UZD	1e8oCD	-18.62
X-ray repair cross-complementing protein 6 (Ku70)	1JEY	3g17AB	-12.32
X-ray repair cross-complementing protein 6 (Ku70)	1JEY	1ry1CD	-14.86
Polo-like kinase 1 (Plk1)	1Q4K	1e8oCD	-3.61
Small Heat shock protein B1 (HSPB1)	3Q9P	1e8oCD	-4.87
Transcription factor E2F1 (E2F1)	2AZE	1e8oCD	-36.94
Glycogen synthase kinase-3 (GSK3B)	1I09	1e8oCD	-8.97
Human Mitogen Activated Protein Kinase (p38 alpha)	3GP0	1e8oCD	-8.15
14-3-3-Protein Zeta (YWHAZ)	1QJA	1e8oCD	-1.73
Serine-Threonine Aurora Kinase A (Ark-1)	1MUO	2r78CD	-10.33



Predicted interactors using p53 DBD conformations from basin C (DNA)	***PDB ID***	***PRISM database Interface***	***Energy (kcal/mol)***
S-Phase kinase associated protein 2 (Skp2)	2ASS	1pxvAC	-2.23
Caspase 3 (Casp-3)	1QX3	1h9sAB	-2.41
RelA-associated inhibitor (iASPP)	2VGE	1ycsAB	-49.25
Apoptosis-stimulating of p53 protein 2 (53BP2)	1YCS	1ycsAB	-72.36
Checkpoint Kinase Chk1 (Chk1)	1NVQ	1e8oCD	-11.87
Cyclin dependent kinase 2 (Cdk2)	1JSU	1jsuBC	-12.63
14-3-3 Protein Gamma (YWHAG)	3UZD	1e8oCD	-0.27
Transcription factor E2F1 (E2F1)	2AZE	1e8oCD	-6.68
Human Mitogen Activated Protein Kinase (p38 alpha)	3GP0	1e8oCD	-12.7
SR Protein Kinase 1 (SRPK1)	1WAK	1h9sAB	-9.9
Dead Box RNA Helicase DDX5 (DDX5)	3FE2	1h9sAB	-1.26
Ras-related protein 1A (RAP1A)	1C1Y	3cgzAB	-9.78

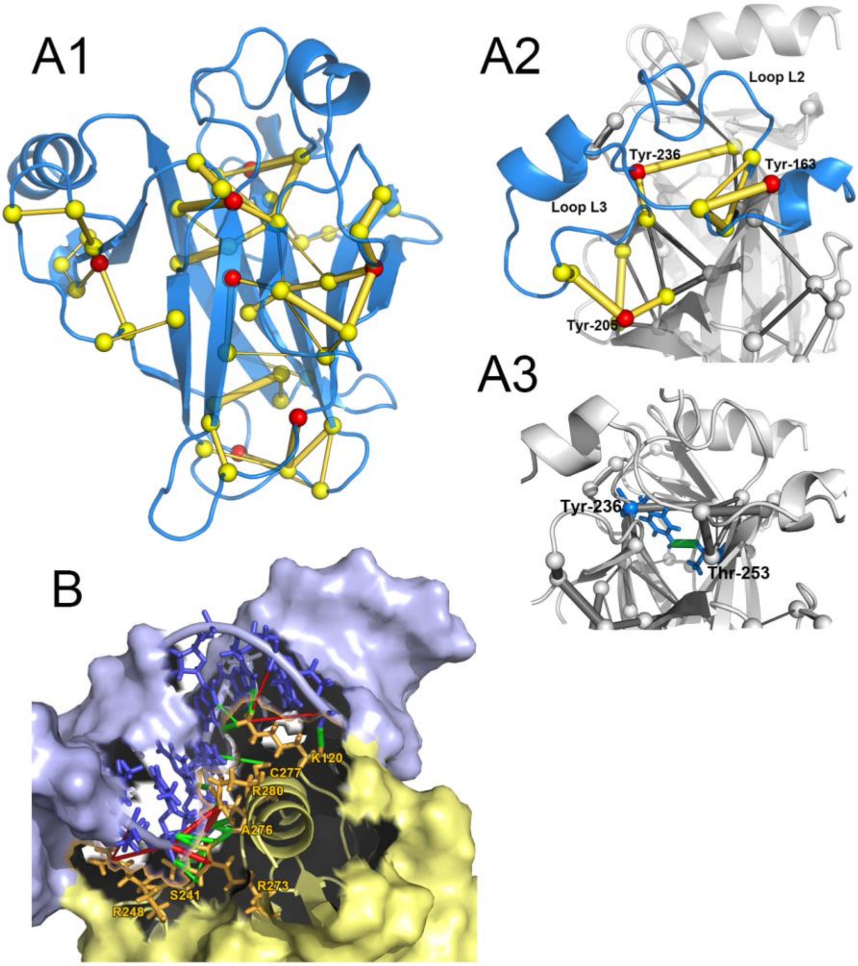


Figure 1S. The p53 DBD and the complex with DNA are stable during the MD simulation time. In order to verify the stability of our p53 DBD and DNA complex during simulation time we performed a detailed analysis of network of the inter- and intra-molecular interactions using a new program developed in our group *PyInteraph* (Tiberti et al., 2014). The hydrophobic interactions generally play a crucial role in the stabilization of the protein core and in the maintenance of thermal stability and 3D structure (Papaleo et al., 2012). Hydrophobic and aromatic residues are usually highly packed inside the protein shielded from the solvent, as in the p53 DNA Binding Domain (DBD) (Cho, 1994). Nevertheless, The p53-DBD has a β-sandwich fold but it is naturally unstable and melts slightly above the body temperatures (42-44°C), becoming prone to be inactivated by oncogenic mutations (Cañadillas et al., 2006). It has been proposed that p53-DBD has evolved to be naturally unstable and this is essential for its activity and regulation (Cañadillas et al., 2006). NMR experiments point out that in p53-DBD several buried polar groups are present, especially of some tyrosine that can be flexible and involved in the formation of sub-optimal H-bond networks, determining its instability (Cañadillas

et al., 2006). We here employed *PyInteraph* to investigate the role of those tyrosine in the stabilization of the p53-DBD, calculating the hydrophobic interactions and their local networks from an MD ensemble and postprocessing the data with *xPyder* (Pasi et al., 2013). **A)** Hydrophobic interactions involved in the core of the p53 DBD We used a significant threshold of persistence of 20%, to discard low populated interactions in the ensemble. The p53 DBD structure is shown in blue cartoon. The Cα atoms of the residues involved in the interactions and of the tyrosines are indicated as yellow and red spheres, respectively. The hydrophobic interactions are represented as cylinders connecting the Cα atoms of residues and their thickness is proportional to the persistence value (A1). Zoom on the interactions that are suggested to stabilize the loop L2 and loop L3 conformations (A2). H-bond between Thr-253 and Tyr-236 (A3). The two residues are highlighted in blue and their side chains are represented as sticks. The H-bond is represented as a green cylinder connecting the Thr-253 Hγ2 and Tyr-236 Oμ atoms. Our analysis shows a wide and highly stable network of hydrophobic interactions (**A)** and points out that Tyr-residues are involved in high persistent interactions in agreement with experimental data that suggest its critical role in the formation of the protein core. Moreover some Tyr-residues (Tyr-163, Tyr-205, Tyr-236) turned out to be localized in key positions in the structure and probably involved in hydrophobic interactions between the β-sheets and loop L2 and loop L3 (**A**). The loops L2 and L3 form the DNA-binding site of p53-DBD and the interactions mediated by the tyrosine were suggested as crucial for the stability of these structural and functionally-relevant elements (Cañadillas et al., 2006). To confirm these results we analysed the H-bonds networks with *PyInteraph* using a distance cutoff of 3.5 Å and map them on the 3D structure. In particular, the analyses show that a highly persistent H-bond (more than 90 %) is present between Thr-253 Hγ2 and Tyr-236 Oμ in our MD ensemble in agreement with most conformers in the deposited NMR ensemble (**A**). Moreover we analysed the MD ensemble of p53 DBD in complex with the DNA and evaluate the maintenance of interactions stabilizing the binding between p53-DBD and DNA by PyInteraph. (B). Hundreds of different p53-response elements (p53-REs) have been identified in the human genome (Wei et al. 2006). They are found in promoters and enhancers associated with the regulation of genes involved in several cellular pathways such as apoptosis and senescence, and they are selectively activated for transcriptional repression or activation by binding to p53. It has been suggested that subtle differences in p53-REs sequence can trigger variances in the interaction patterns and induce allosteric alterations on p53 DBD, which can, in turn, affect the recruitment of coregulator and the organization of tetrameric p53 in order to activate specific functions (Pan et al., 2010). A cutoff of 20% was identified as a significant threshold of interaction persistence. **B)** The p53-DBD and the DNA are shown in yellow and light blue and represented as cartoon and surface. The salt bridges and H-bonds between the protein residues and DNA are shown as red and green cylinders, respectively, with thickness proportional to the persistence of the interaction in MD ensemble. The residues involved in the interactions are highlighted with sticks.Our analyses allow the description of all the intermolecular relevant electrostatic interactions between p53 DBD and DNA, in a MD framework, showing an overall agreement with a previous study. In particular, we identified highly persistent and sequence-specific contacts with the DNA major groove by residues in H2 helix and loop L1 of p53 DBD, as well as contacts with the DNA minor groove by residues in loop L3 (B). Moreover, salt bridges between the DNA-backbone phosphate groups and the protein can be identified (B). In particular, residues relevant for the interaction with the DNA major groove are K120, C277, and R280. K120 also interacts with the phosphate groups of Gua7 and Gua8. R280 is involved in salt-bridges and H-bonds with phosphates of Gua10’ and Thy11’. The interactions with the DNA minor groove are mostly mediated by R248 in loop L3, A276 backbone amide, R273, and S241.

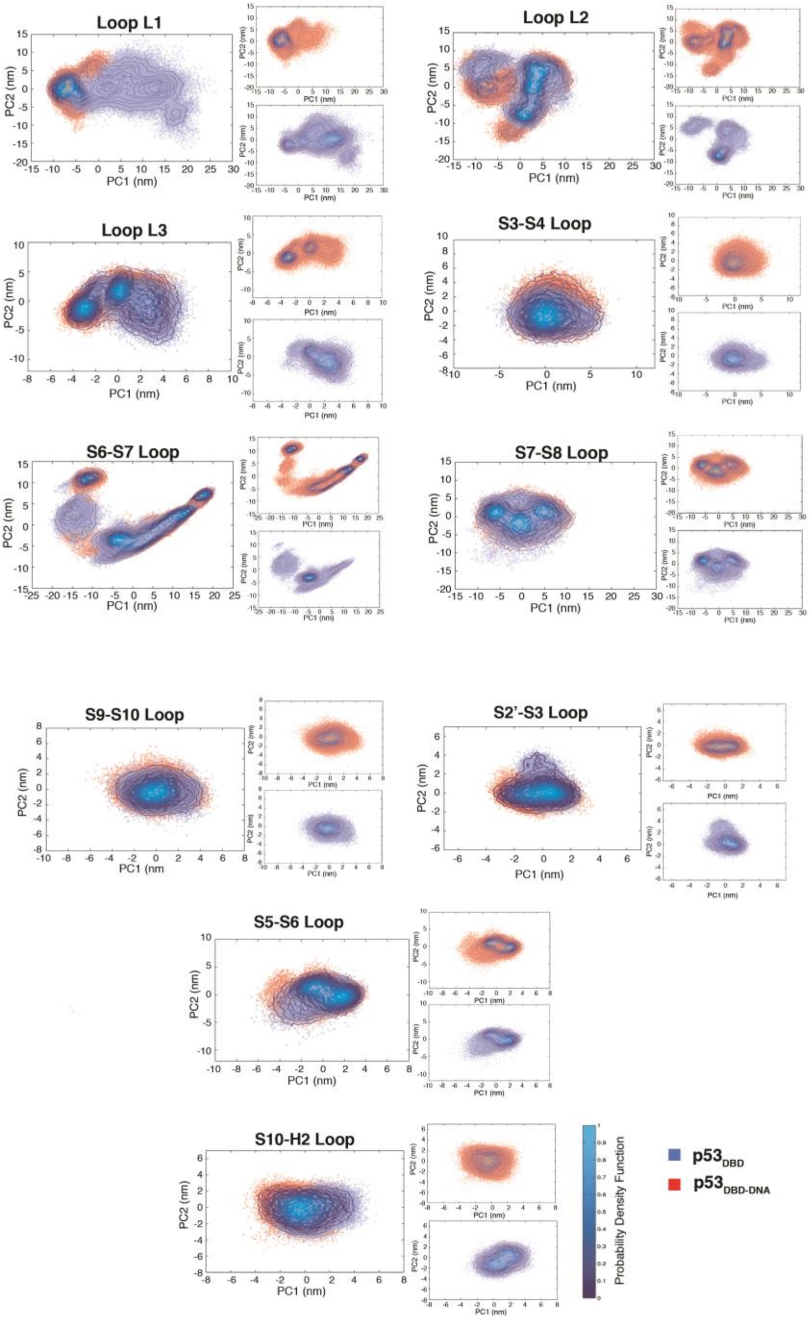

Loop L1
Loop L2
Loop L3
S3-S4 Loop
S6-S7 Loop
S7-S8 Loop
S9-S10 Loop
S2'-S3 Loop
S5-S6 Loop
S10-H2 Loop
PC1 (nm)
PC2 (nm)
PC1 (nm
Probability Density Function
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
p53$_{DBD}$
p53$_{DBD-DNA}$

Figure 2S. Projections of the simulation frames along the first two principal components of the essential subspace. Two-dimensional (2D) projections along the first and second principal components (PCs) derived by principal component analysis (PCA) of a concatenated trajectory, which include all the MD replicates of both p53$_{\text{DBD}}$ and p53$_{\text{DBD-DNA.}}$ We carried out the covariance matrix for PCA calculation for each loop of the p53 DBD upon fitting on the Cα atoms of the β-sheet core. The simulation frames from p53$_{\text{DBD}}$ and p53$_{\text{DBD-DNA}}$ are highlighted in blue and red, respectively. The majority of the loops are insensitive to DNA-interaction, whereas we observed differences in the subspace sampled in presence or absence of DNA for loops L1, L3, and loop S6-S7. In the small inserts on the left side of each plot, the values calculated from only p53$_{\text{DBD}}$ (blue) and only p53$_{\text{DBD-DNA}}$ (red) are shown.

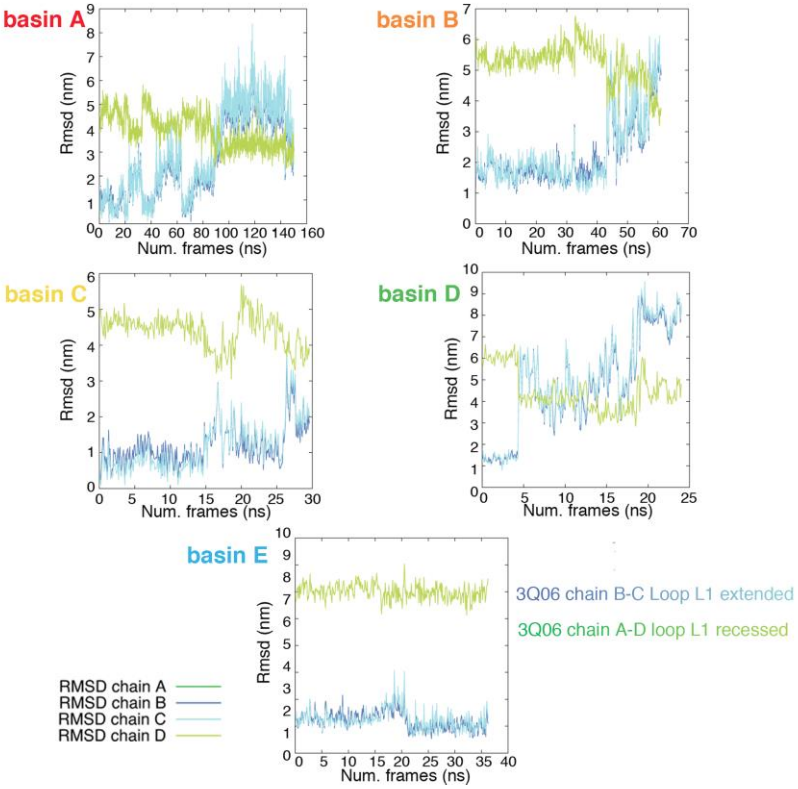


Figure 3S. Conformations of loop L1 in the DNA-bound and –unbound p53 DBD simulations. All-atoms root mean square deviation (rmsd) profiles of loop L1 residues (114-124) in the structures extracted from each of the PCA regions (*A,B,C,D,E*) with respect to the structure that p53 DBD has in each of the monomer (chains A, B, C, D) of the X-ray tetrameric structure in complex with DNA (PDB entry: 3Q06). The monomers in chain B and C of the tetrameric X-ray structure are characterized by extended conformations of loop L1, whereas the monomers in chain A and D feature recessed conformations of loop L1. We calculated the rmsd upon fitting on the Cα atoms of the β-sheet core of p53 DBD for each single chain in the tetramer.

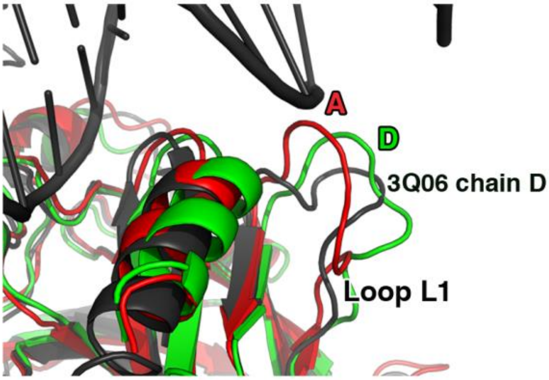


Figure 4S. Comparison of *A*-like and *D*-like states from our simulations with the recessed conformation of loop L1. Here, we show the comparison between the conformations that loop L1 has in the average structure extracted from the *A* and *D* PCA regions and the recessed conformation observed in the crystallographic structure (PDB entry 3Q06, chain D).

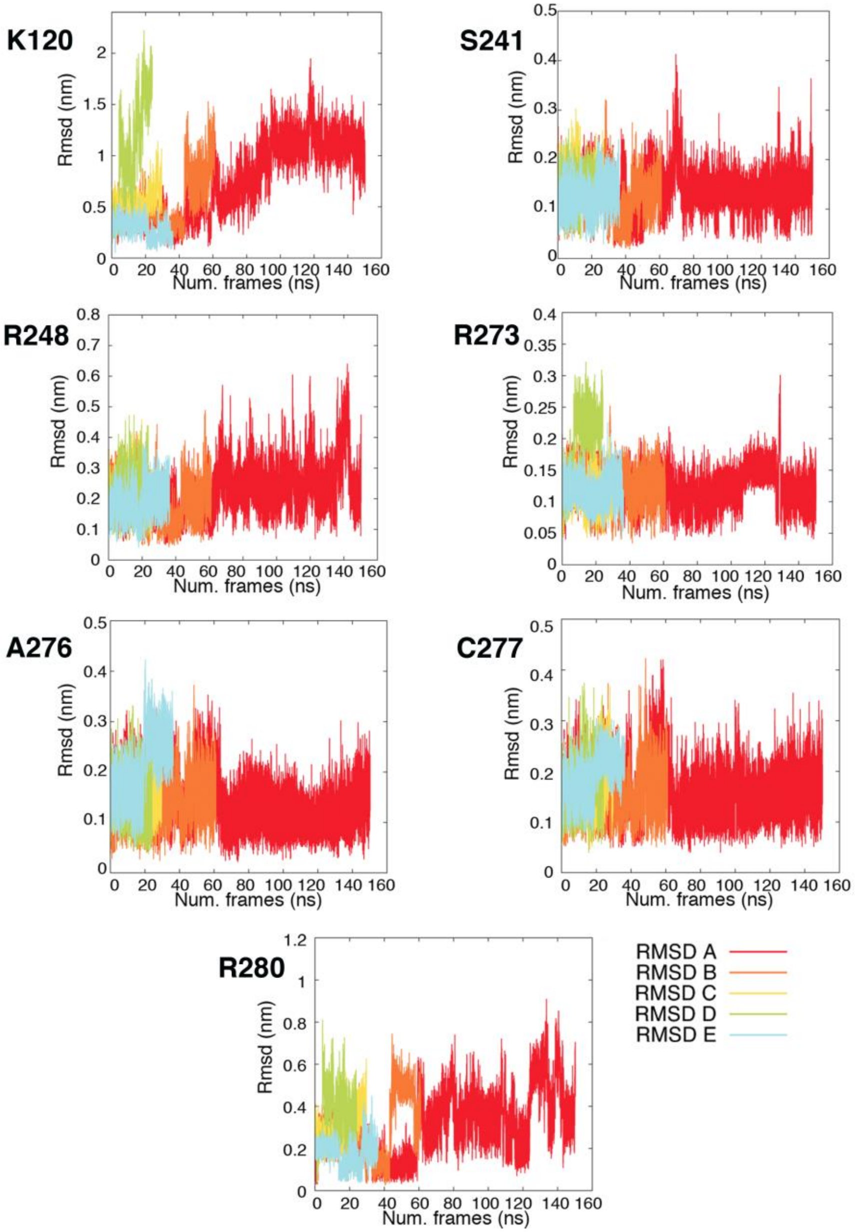


Figure 5S. Root mean square deviation of p53 DBD residues that are at the DNA binding interface in the p53$_{\mathrm{DBD}}$ simulations. We calculated the all-atom rmsd of p53 residues at the interface for DNA binding using as reference the initial structure upon fitting on the Cα atoms of the β-sheet core of p53 DBD. We calculated the rmsd separately for each region of the 2D PCA subspace illustrated in Figure 2S.

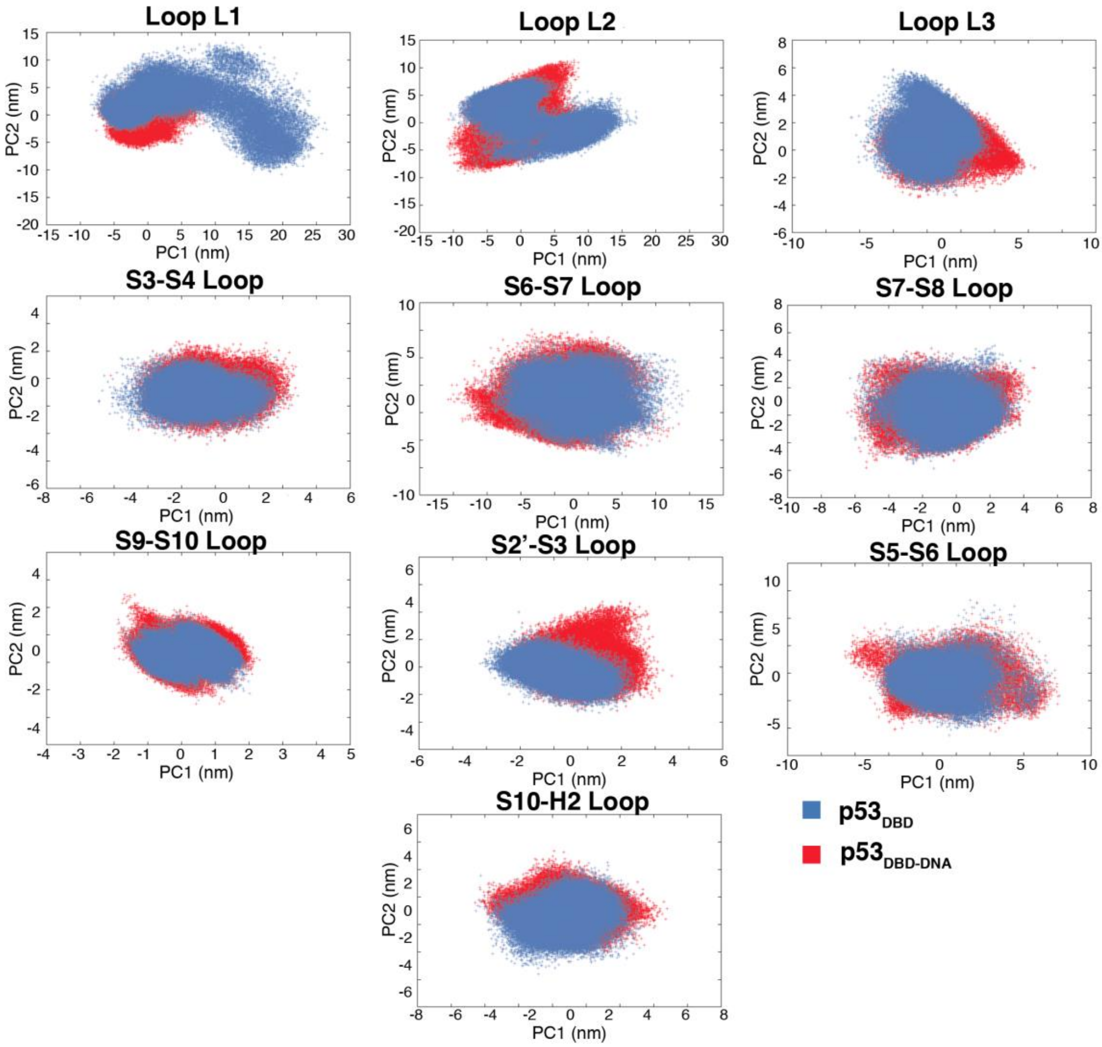


Figure 6S. Projections of the simulation frames along the first two principal components of the essential subspace. Two-dimensional projections along the first and second principal components (PCs) derived by principal component analysis (PCA) of a concatenated trajectory, which include all the MD replicates of both p53$_{DBD(91-289)}$ and p53$_{DBD(91-289)-DNA.}$ We carried out the covariance matrix for PCA calculation for each loop of the p53 DBD upon fitting on the Cα atoms of the β-sheet core. The simulation frames from p53$_{DBD(91-289)}$ and p53$_{DBD(91-289)-DNA}$ are highlighted in blue and red, respectively. The majority of the loops are insensitive to DNA-interaction, whereas we observed differences in the subspace sampled in presence or absence of DNA for loops L1.

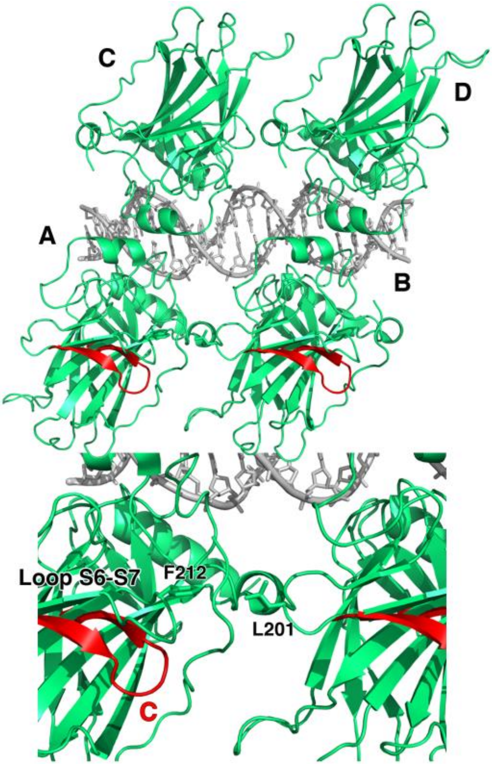


Figure 7S. Comparison of conformations of loop S6-S7 from our MD simulations with the structure of the loop in the tetrameric assembly of p53 DBD. The tetramer of DNA-bound p53 is composed by two dimers of p53 DBD interacting each other (monomer A interacts with monomer C and B with D). The interaction interface involves helix H1, the Zn^{2+} binding region and parts of loops L2 and L3. Moreover, the dimers pack themselves side to side to form a tetramer. Indeed, in a recent crystallographic structure (PDB entry 3KMD (Chen et al., 2010), green cartoon), the two dimers (chains A and C that interact with chains B and D) interact with the disordered N-terminal region, loop L2 of one molecule (A and D) and the loop S5-S6, loop S7-S8 and the βstrands S3 and S8 from the other molecule (B and C). Loop S6-S7 is solvent-exposed in all the monomers with the only exception of Phe212 that can interact with Leu201 from another subunit. Here, we used the average structure of the *C* state from our simulations as a reference to illustrate the conformation of loop S6-S7 (residues 202-216), which is highlighted in red. We superimposed the *C* average structure to the experimental crystallographic structure using as a reference the Cα atoms of the core β-sheet. The Phe212 of loop S6-S7 and the Leu201 of S5-S6 loop, which are involved in the interaction between the adjacent dimers of the experimental tetrameric structure, are showed as sticks. Moreover, we analysed other experimental deposition for the p53 tetramer, which are also supported by recent NMR and Electron Microscopy experiments (Bista et al., 2012, Melero et al., 2011) using the following PDB entries: 1TSR, 1TUP, 2GEQ, 2AC0, 3KZ8, 3KMD (Chen et al., 2010, Malecka et al., 2009, Kitayner et al., 2006). We noticed that all the conformations of loop S6-S7 sampled in our MD simulations can fit in the quaternary complexes proposed for p53 DBD tetramerization without causing any clash.

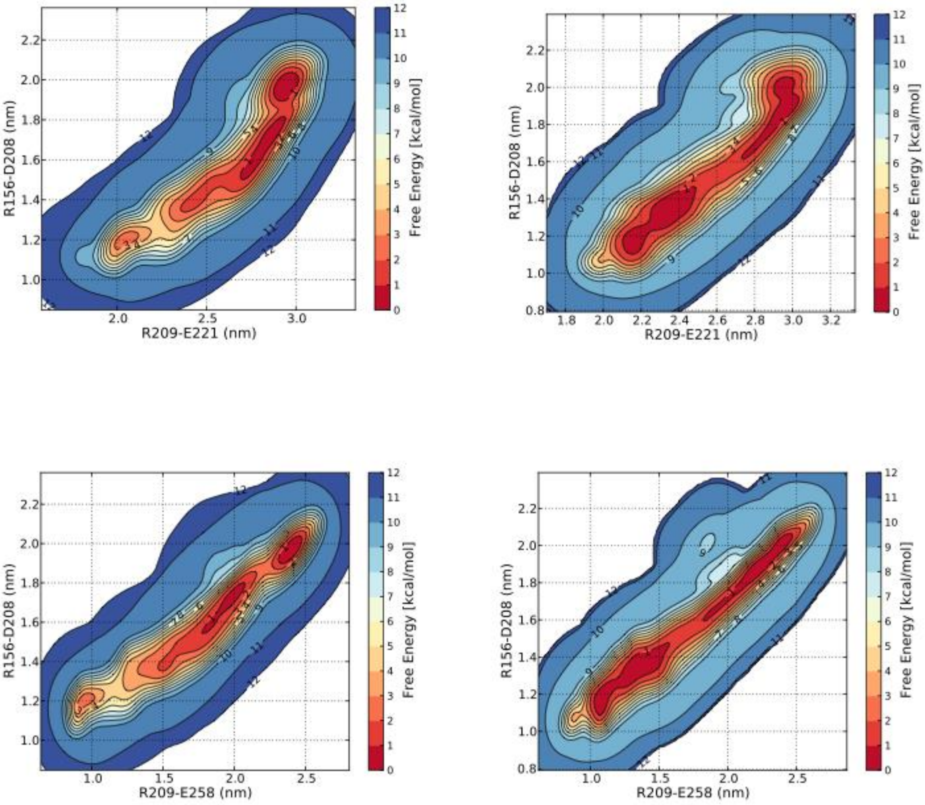


Figure 8S. Additional two-dimensional FES profiles from PT-metaD simulations. The two-dimensional FES profiles are shown for the additional collective variables that are not shown in the main text. p53$_{DBD(91-289)}$ and p53$_{DBD-DNA(91-289)}$ are shown on the left and right panels, respectively. It is possible to observe that *C*-like states (the minima corresponding at lower distances in the plots) are only a minor population in the free state, whereas they are more than four fold populated when DNA is bound.

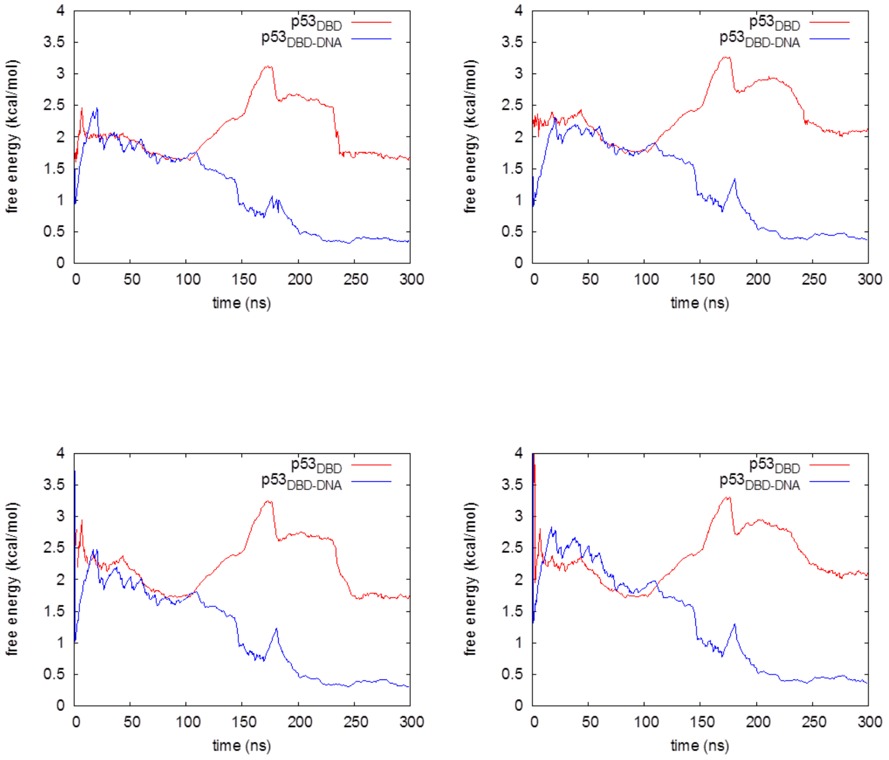


Figure 9S. Convergence of the PT-metaD simulations of $p53_{DBD(91\text{-}289)}$ and $p53_{DBD\text{-}DNA(91\text{-}289)}$. The plots show the free-energy difference between *A*-like and *C*-like states in absence and presence of DNA for each collective variable over the simulation time. For each time point, the plots show the free energy difference calculated using the simulations up until that time. The mono-dimensional plots of each collective variable (CV) were used for this calculation and the distances which reflect the *C*-like (and *B*-like) states were defined as all the states at a distance lower than 1.45, 1.4, 2.5 and 1.5 nm for CV 1, 2,3 and 4, respectively, in agreement with what observed also in the unbiased MD simulations.

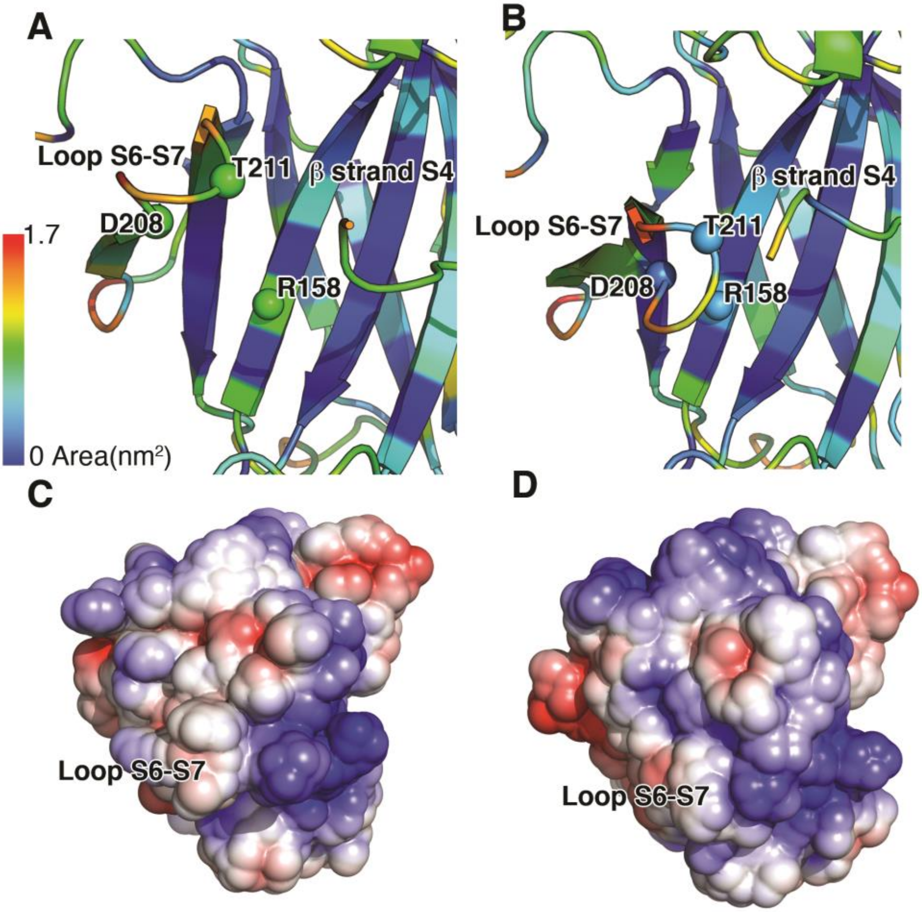


Figure 10S. Solvent accessible surface and electrostatic potential surface. We performed the analysis using as references the average structure from basin *A* (panels A and C) and the average structure from basin *C* (panels B and D). We observed that conformational changes in loop S6-S7 from *A*-like to *C*-like conformations alter locally the solvent accessible surface (SAS) of the p53 DBD, as shown in panels A and B for p53$_{\mathrm{DBD}}$ and p53$_{\mathrm{DBD\text{-}DNA}}$, respectively. In particular, *C*-like DNA-bound states are characterized by less accessible residues in the β-strand S4, such as Arg158, and in loop S6-S7 such as Thr211 and Asp208 compared to the *A*-like DNA-unbound states. We indicated the Cα atoms of these residues with spheres. The average SAS values of each residue are highlighted as color gradient from low (dark blue) to high (red) accessibility. The electrostatic potential surface calculated by APBS (Baker et al., 2001) is shown in panels C and D for p53$_{\mathrm{DBD}}$ and p53$_{\mathrm{DBD\text{-}DNA}}$, respectively. The electrostatic potential maps are projected on the solvent accessible surface of the p53 DBD. The negatively and positively charged groups are colored in red and blue, respectively, with the intensity of the color proportional to the local potential.

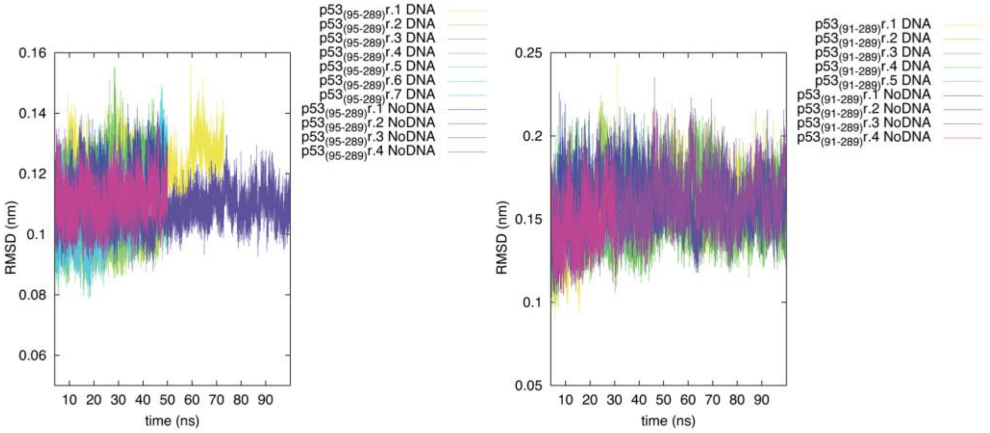
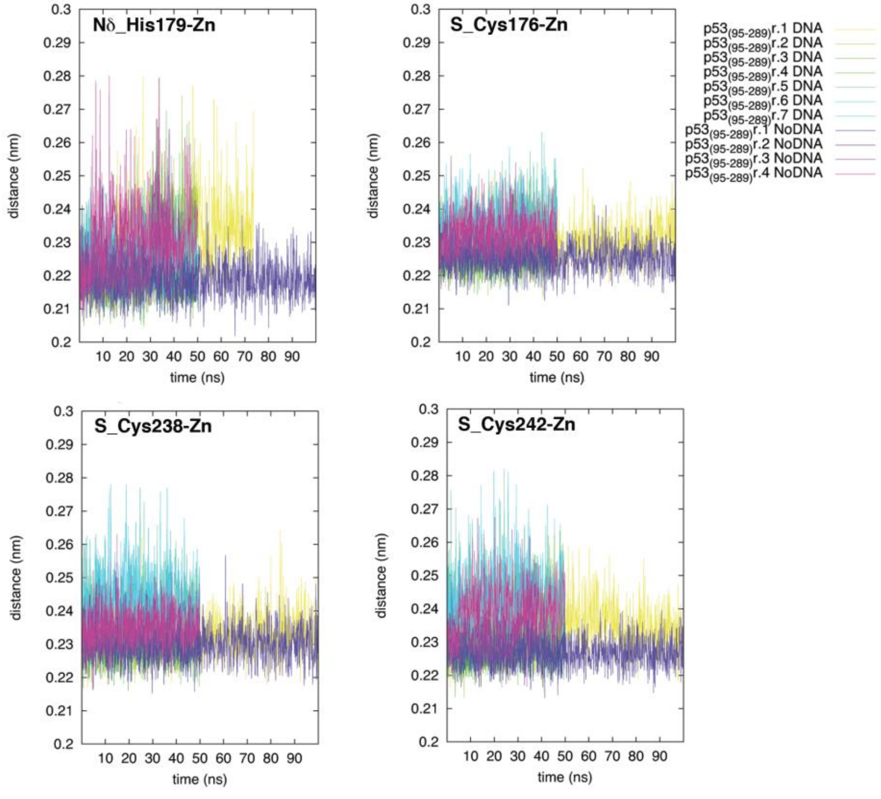


Figure 11S. Main chain rmsd profiles of our unbiased MD simulations. We calculated the main chain rmsd over the simulation time using as a reference the initial crystallographic structure, considering only the secondary structural elements, i.e. α-helices and the β-sheet core.



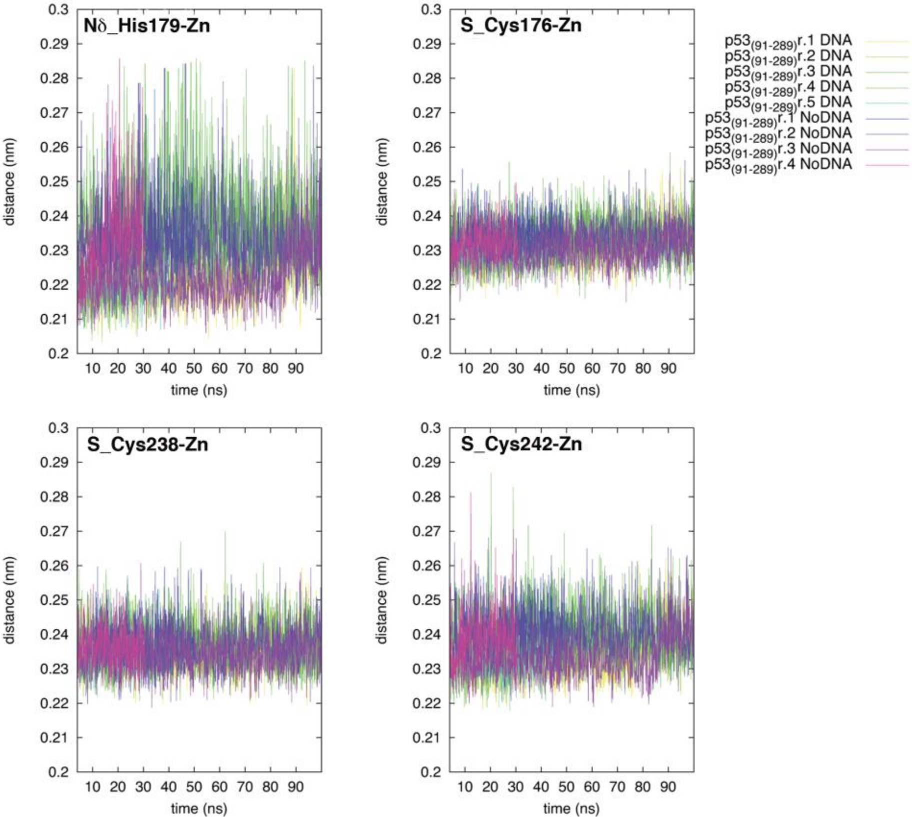


Figure 12S. Distances between Zn^{2+} ion and its coordinating residues (Cys176, Hys179, Cys238 and Cys242) over the simulation time in our unbaiased MD simulations.

References

Anderson C, Appella E (2009) Signaling to the p53 tumor suppressor through pathways activated by genotoxic and non-genotoxic stresses. *Handb Cell Signal* 264:2185–2203.

Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A* 98:10037–41.

Bista M, Freund SM, Fersht AR (2012) Domain-domain interactions in full-length p53 and a specific DNA complex probed by methyl NMR spectroscopy. *Proc Natl Acad Sci U S A* 109:15752–6.

Bussi G, Donadio D, Parrinello M (2007) Canonical sampling through velocity rescaling. *J Chem Phys* 126:014101.

Canadillas, J. M., H. Tidow, S. M. Freund, T. J. Rutherford, H. C. Ang and A. R. Fersht (2006). Solution structure of p53 core domain: structural basis for its instability. *Proc Natl Acad Sci U S A* 103(7): 2109-2114.

Carmena M, Earnshaw WC (2003) The cellular geography of aurora kinases. *Nat Rev Mol Cell Biol* 4:842–54.

Chen Y, Dey R, Chen L (2010) Crystal structure of the p53 core domain bound to a full consensus site as a self-assembled tetramer. *Structure* 18:246–56.

Darden T, York D, Pedersen L (1993) Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J Chem Phys* 98:10089.

Fraser J a, Vojtesek B, Hupp TR (2010) A novel p53 phosphorylation site within the MDM2 ubiquitination signal: I. phosphorylation at SER269 in vivo is linked to inactivation of p53 function. *J Biol Chem* 285:37762–72.

Hess B, Bekker H, Berendsen H, Fraaije J (1993) LINCS: A linear constraint solver for molecular simulations., *J Comput Chem*, 12: 1463-1472.

Kitayner M et al. (2006) Structural basis of DNA recognition by p53 tetramers. *Mol Cell* 22:741–53.

Malecka KA, Ho WC, Marmorstein R (2009) Crystal structure of a p53 core tetramer bound to DNA. *Oncogene* 28:325–33.

Melero R et al. (2011) Electron microscopy studies on the quaternary structure of p53 reveal different binding modes for p53 tetramers in complex with DNA. *Proc Natl Acad Sci U S A* 108:557–62.

Pan, Y. and R. Nussinov (2010). Lysine120 interactions with p53 response elements can allosterically direct p53 organization. *PLoS Comput Biol* 6(8).

Petty TJ et al. (2011) An induced fit mechanism regulates p53 DNA binding kinetics to confer sequence specificity. *EMBO J* 30:2167–76.

Qiyuan L., S. Kaneko, L. Yang, R. I. Feldman, S. V. Nicosia, J. Chen, J. Q. Cheng (2004) Aurora-A Abrogation of p53 DNA Binding and Transactivation Activity by Phosphorylation of Serine 215. *J Biol Chem* 279 (50): 52175–52182.

Tiberti, M., G. Invernizzi, M. Lambrughi, Y. Inbar, G. Schreiber and E. Papaleo (2014). PyInteraph: a framework for the analysis of interaction networks in structural ensembles of proteins. *J Chem Inf Model* 54(5): 1537-1551.

Warnock LJ, Raines SA, Milner J (2011) Aurora A mediates cross-talk between N- and C-terminal post-translational modifications of p53. *Cancer Biol Ther* 12:1059–68.

3.2 Computational characterization of the role of metals coordination in the dynamics of proteins

3.2.1 Introduction

Metals are essential for life, since at least one third of known proteins in eukaryote contain metals to carry out their biological function. Binding of metals ions serves essential biological processes such as enzyme catalysis, cellular signaling, control of oxidative stress, and structural stabilization. In particular zinc, the second most abundant metal cation in human body, has a dominant importance in proteins both for catalysis and structural functions (Andreini et al., 2006, Dudev and Lim, 2008). The requirement of a strictly regulated homeostasis of zinc is emphasized by the observation that modifications of its physiological levels in cells, such as after the intoxication with non-essential metals like cadmium (Makowski and Sunderman, 1992, Hartwig, 2010) determine an alteration of its roles and functions in proteins and are involved in the development of several pathological states, like cancer and neurodegenerative diseases (Loh, 2010, Breydo and Uversky, 2011, Xu et al. 2011). In fact cadmium is a serious environmental pollutant and human carcinogen (Hartwig, 2010, Thevenod, 2010). Main exposures at cadmium occur at workplaces but also in the general population, mainly through emissions in the ambient air, food and industrial products such as batteries, solder, plastics, paints, alloys and tobacco smoke. Cadmium has a very long emivite (>25 years) and tends to accumulate in human cells after long term exposure, principally in liver, kidneys and lungs, inducing the development of different types of cancer (Hartwig, 2010). Cadmium exerts a complex action on the cell determining alteration of the DNA damage response and repair processes, inhibition of apoptosis, alteration of tumor suppression functions, deregulation of gene expression and cell cycle control, increase the level of reactive oxygen species (ROS), determining a severe global genomic instability, associated with resistance to apoptosis, induction of cellular hyperproliferation, hypermutability and accumulation of carcinogenic mutations. The impact of cadmium substitution and its principal effects occur through interferences with zinc in multiple Zn-binding proteins, like many transcription factors and several cell cycle control and DNA repair proteins, destabilizing the protein structures, inhibiting functions and leading to misfolding and aggregations processes in an still unknown molecular mechanism (Kuwahara and Coleman, 1990, Hanas and Gunn 1996, Roesijadi et al., 1998, Waisberg et al., 2003, Huang et al., 2004, Martelli et al., 2006). Exposure to cadmium inhibits proteins such as mammalian xeroderma pigmentosum group A protein (XPA) poly-ADP-ribose polymerase 1, SP1 and tumor suppressor p53 (Hartwig et al., 2002). Moreover it has been proposed that Cd^{2+} ions can directly replace Zn^{2+} ions in protein coordination sites and this is the leading molecular mechanism of cadmium toxicity (Hartwig, 2001, Martelli et al., 2006). According to Pearson's classification the large and easily polarized cation

Cd^{2+} is a soft Lewis acid, whereas Zn^{2+} is borderline. Therefore, cadmium can combine with high affinity to soft ligands, such as the sulfur atoms of the cysteine residues, possibly displacing Zn from its coordination sphere in Zn-binding proteins. This hypothesis is corroborated by the fact that cadmium and zinc have similar chemical-physical properties, are in the same group of the periodic table and have a common oxidation state. Only a few in vitro and in vivo experimental evidences have been reported but they pointed out important evidence in favor of this mechanism, such as the structural investigations performed on XPA (Buchko et al., 2000, Kopera et al., 2004). Moreover the structure of the SUPERMAN (SUP) protein zinc-finger domain resolved by NMR techniques, both in its native form and with zinc replaced by cadmium, supports this hypothesis, and gives information on the structural effects induced by cadmium (Isernia et al., 2003, Malgieri et al., 2011). In this context, the exact mechanism and physicochemical principles governing protein-metal relationship remain elusive. Little is known about the cadmium interferences with metal binding and how it is associated with diseases. The structural characterization of Zn-binding proteins, and the investigation of the mechanisms of metal-mediated conformational changes, is essential in order to understand the role of metals in these biological relevant problems. Since the lacking of robust experimental evidence, and due to the complexity of the problem, an in silico approach is essential in order to obtain a characterization of the conformational landscape and investigate dynamical and structural properties of metal binding proteins at atomic level. In this context, recent evidences showed that with the increasing accuracy and efficiency of force fields, atomistic explicit solvent Molecular Dynamics (MD) simulations and enhanced sampling MD techniques can correctly describe complex events in proteins (Klepeis et al., 2009, Lindorff-Larsen et al., 2012, Papaleo et al., 2012). Several investigations proposed that the MD can be valuable to describe Zn (Dudev and Lim, 2008, Hu and Ryde, 2011, Chakravorty et al., 2012, Seneque and Latour, 2010, Wise-Scira et al., 2012, Miller et al., 2010), but approximations are still present in the force field in the modelling of coordination with metals, especially when long timescales are simulated (Calimet and Simonson 2006). The description of cadmium is also very critical. Overall these studies all point out the necessity of improved force-field parameters to describe metal-binding proteins (Hu et al., 2011). Two main types of approaches have been implemented in classical pairwise ff: nonbonded and bonded models. In the nonbonded model, the interaction between metal and its ligands is represented by van der Waals and electrostatic terms (Stote and Karplus, 1995, Donini and Kollman, 2000, Calimet and Simonson, 2006, Sakharov and Lim, 2009). This model is used for its simplicity and efficiency in the investigation of structure and dynamics of zinc containing proteins but it has approximations in the treatment of strong electrostatic interactions, with overly strong propensity of zinc to get close to negatively charged aminoacids, and in the maintenance of metal binding site that may distort, loosing correct coordination. In the bonded model, artificial bonds are introduced into the potential energy function to model the binding between metal ion and its ligands using explicit harmonic energy terms (Toba et al., 1999, Zhang et al., 2004, Calimet and Simonson, 2006). The main property of bonded models is that the metal binding structure is preserved by the nature of the energy function. Also a semi-bonded model using tetrahedral charge dummy atoms around zinc was proposed. Previous studies showed that, using bonded models, the tetra-coordination of zinc binding group and residues in its first coordination shell can be maintained during simulation and permitted realistic MD simulations (Pang, 2001) but they caused artifact in conformational sampling linked to the over freezing of the binding site. Associated to these approaches, different methods were used to take into account effects like structure-dependent polarization and the charge transfer effects, in order to accurate investigate the long-range electrostatic interactions, that have been pointed out to be important for metalloprotein (Russell and Fersht 1987). Simple approaches used nonbonded model and consider the polarization and charge transfer effect through parameterization of ff based on high-level Quantum Mechanics (QM) calculations (Ryde, 1995, Calimet and Simonson, 2006) or by a model that introduces distance-dependent partial charges on zinc (CTPOL) (Sakharov and Lim, 2009). Complex polarizable potentials and forcefield like SIBFA, AMOEBA, APPC, QPCT were developed for zinc binding protein and they can give very accurate ion−ligand interaction energies and structures, but there are still problems and tendency to over constrain the system. On the other side the Quantum Mechanics (QM) and combined QM/MM methods have been applied to investigate zinc-binding proteins (Elstner et al., 2003). However, due to heavy computational cost, these methods are severely limited in their applications to large protein systems or long time simulation and to investigate conformational alterations such as induced by cadmium. In this context we developed a new approach for parameterization of MD classical force field based on molecular mechanics (MM) and quantum chemical calculations at the density functional theory (DFT) level, that permits to develop highly optimized parameters for the metal ions, that can be readily used in MD simulations to accurately describe the coordination of metals in metal-binding proteins, permitting to describe in details their structural properties and dynamics. QM calculations were performed in collaboration with Prof. Maurizio Bruschi at the Department of Environmental and Earth Sciences at the University of Milano-Bicocca. QM calculations pointed out that binding energy profoundly changes between different metal coordination sites suggesting that each coordination mode has to be treated with specifically optimized parameters. We developed optimized force-field parameters from CHARMM22/CMAP (Mackerell et al., 2004) for zinc ions and residues involved in the metal binding site for a Cys$_{s}$His$_{s}$ coordination site, taking into account polarization and charge transfer effects and the covalent character of the metal-ligand bond. We also developed new force-field parameters for cadmium that were missing in the CHARMM22/CMAP. In the present project we performed a preliminary analysis of our approach and examined the behavior of our optimized parameters carrying out atomistic explicit-solvent MD simulations on the zinc finger domain of SUP protein and compared our results with the experimental structures of SUP, pointing out the effectiveness of our new approach. We shown that force field based only on nonbonded interaction to treat the coordination with the metal ions in proteins leads to poor accuracy for the association energies and for the binding site geometries. The addiction of binding terms and the use of our optimized force field permits to sample structures that are close to the experimental ones, while non-optimized parameters lead to poorer structures. Our approach is simple but straightforward an our preliminary results suggest that is an effective tool to study by MD simulations metal binding proteins in solution. We are now further testing our new force-field parameters collecting microsecond and enhanced-sampling MD simulations. In the future we will develop optimized force-field parameters also for other types of metal coordination site.

3.2.2 Methods

Ab initio calculations

A first set of Quantum Mechanics (QM) *ab initio* calculations were performed on zinc model complex, resembling the coordination site present in the SUPERMAN protein zinc finger domain, that is composed by two cysteines and two histidines $Zn(Cys)_2(His)_2$. The model complex is defined by including the atoms of Cys and His residues side chains till the Cα, thus using as ligands two ethylimidazoles (EtImid) and two ethylthiolates (SEt), so that the general formula is $[Zn(SEt)_2(EtImid)_2]$. For comparison we also performed calculations for other $[Zn(SEt)_y(EtImid)_x]$ (y=0-4; x=4-y) model complexes by varying the number of ethylthiolates and ethylimidazoles coordinated to the metal atom (data not shown). A second set of calculations was performed on the model complex $[Cd(SEt)_2(EtImid)_2]$ obtained by replacing Zn with Cd to the $[Zn(SEt)_2(EtImid)_2]$ model complex. In the protein, histidine residues are coordinated to metals with the δ–nitrogen of their imidazole ring, and model complexes were generated accordingly.

Geometry optimizations and energy calculations were carried out in the framework of the Density Functional Theory (DFT) (Parr et al., 1989) by using the three parameter hybrid exchange correlation functional B3LYP (Becke, 1993, Lee et al., 1988) and the triple-ζ quality basis set TZVP (Schafer, et al., 1994). All calculations were performed with the Turbomole suite of programs (Schafer et al., 1992). Atomic charges were calculated using the Natural Bond Orbital (NBO) approach (Reed et al., 1988). Moreover for each system potential energy curves were calculated for the dissociation of each ligand from the metal ion, performing single point energy calculation at DFT level with B3LYP/TZVP basis sets. Each ligand was individually separate at several distances between metal ion- nitrogen δ or metal ion-thiol sulfur (from 1.8 Å to 5.0 Å then at 10.0 and 100.0 Å) and then the energy of the system was calculated. We also performed the DFT calculations using an implicit solvent model that was applied to mimic the protein interior using an organic solvent environment, as suggested by Duan et al. ($\varepsilon = 4$) and to avert overpolarization. The results are very similar to those of the previous calculation and they will not be further taken into account.

.

Force- field parameter optimization and fitting

Force-field parameters were derived for the Zn $(Cys)_2(His)_2$ and $Cd(Cys)_2(His)_2$ coordination mode with the two cysteines modeled as ethylthiolates and two histidines modeled as ethylimidazoles, in order to fit the geometry of minimum energy obtained by DFT calculations. In our new force field the zinc and cadmium–ligand interactions were modelled with harmonic, bonded terms, which prevent the ligands from leaving the coordination sphere but at the same time the terms do not excessively constrain the interaction. Moreover additional angle and dihedral terms was introduced in the force field to better describe the geometry of the metal site (see Table 2 for the definition of force-field terms). We developed a new tool that performs the optimization of the parameters present in the CHARMM22/CMAP force field based on the QM results. The program performs geometry optimization and single point energy calculations with classical molecular mechanics using the program NAMD and the CHARMM22/CMAP with the new terms. To optimize the force-field parameters and reproduce QM results the program uses a simulated annealing approach minimizing an error function composed by the root mean square (rms) deviation between the MM and QM optimized geometries of the zinc and its four ligands and differences in the potential energies of the dissociation curves obtained by MM and QM calculations. Parameters were accepted when the rms deviation error value was under a chosen threshold of 0.01 Å and energy error over the dissociation potential energy curves was under 0.1 kcal/mol. For the MM calculations atomic partial charges were taken directly from the DFT calculations on small tetrahedral complexes for Zn^{2+} and Cd^{2+}. NBO charges were used and this choice is reasonably consistent with the development of the CHARMM ff. In addition in order to take into account the different polarization effects due to the specific binding site of the metal atom, we build up a system than redistribuites residual charges (net charge of the system minus the QM charge of the metals) starting from the charge in the CHARMM22/CMAP, over a set of heavy atoms in histidine (Nδ, Cδ, Nε, Cε, Cγ) and cysteine (S, Cβ, Cα) on the basis of the distances of each atoms from the metal ions. The optimized parameters are shown in table 1. All other parameters were taken from the CHARMM22/CMAP.

Molecular dynamics protocol

MD simulations were performed for two structure of SUPERMAN protein zinc-finger domain (SUP-zf) in explicit solvent. The SUP structures were the wild type NMR structure solved with zinc (SUP-zf-Zn, PDB entry 1NJQ, residue 42-78, Isernia et al., 2003) and the NMR structure solved with zinc replaced by cadmium, (SUP-zf-Cd, PDB entry 2L1O, residue 42-78, Malgieri et al., 2011). The initial structures of both Sup-zf-Zn and SUP-zf-Cd were model 1 out of the 20 models in each of the NMR ensemble, these models were chosen because they are closest to the structure obtained by averaging over all the models. The systems were solvated in a dodecahedral box (minimum distance between protein and box edges: 1 nm) of TIP3P water molecules at 150 mM NaCl, applying periodic boundary conditions. We used the CHARMM22/CMAP force field (Mackerell et al., 2004) for the proteins, except for the force-field parameters newly parameterized in the present work. Each system was initially relaxed by 10000 steps of energy minimization by the steepest descent method. The optimization step was followed by 50 ps of solvent equilibration at 300K, while restraining the protein atomic positions using a harmonic potential. Each system was then equilibrated to the target temperature (300 K) and pressure (1 bar) through thermalization and a series of pressurization simulations of 100 ps each. We performed productive MD simulations using LINCS algorithm (Hess et al., 1997) to constrain heavy-atom bonds, allowing for a 2 fs time-step. Long-range electrostatic interactions were described by the Particle-mesh Ewald summation scheme.Van der Waals and Coulomb interactions were truncated at 0.9 nm. We collected overall 5 independent MD simulations (replicates) of 100 ns each (Table 1S) at 300 K and 1 atm using the isothermal-isobaric ensemble (NPT) and an external Berendsen bath with thermal and pressure coupling of 0.1 and 1 ps, respectively. The non-bonded interaction list was updated every 10 steps and conformations stored every 4 ps in the productive MD runs. In all the simulations, the cysteines coordinating the zinc were assumed deprotonated (all-thiolate model) consistent with the data present in literature for zinc-finger domain of other proteins and as demonstrated by the symmetry in the zinc coordination in SUP structure, with all the Zn–S distances within 2.27 and 2.32 Å, also in agreement with the optimized geometry from QM calculations.

Simulations analysis

Intramolecular interactions and their networks, electrostatic, aromatic and hydrophobic interactions, were analyzed in details by Pynteraph, a new program based on graph theory developed by our group (Tiberti et al., 2014) employing a persistence cut-off of 20% and a distance cut-off of 0.5 nm. Hydrogen bonds were analyzed using a persistence cut-off of 20%, a distance cut-off between donor and acceptor group of 0.3 nm and a minimum donor–H-acceptor angle of 120°. Interactions networks were analyzed with xPyder (Pasi et al., 2012). Pynteraph and Xpyder represent pairwise relationships related to protein structures, such as intramolecular interaction networks, as two-dimensional matrices. A network is described as a set of points (nodes) and connections between them (edges). A path is defined as a sequence of nodes for which an edge always exists between two consecutive nodes of the path. A matrix describing the persistence of each class of interactions was used as an input file. The program represents each residue of the matrix as a node of a simple, weighted graph connected by edges, whose weights are defined by the persistence of the interaction in the MD ensemble (i.e. the number of trajectory frames in which the interaction was present, over the total number of frames).

3.2.3 Results

It is essential to to develop computational methods especially molecular dynamics (MD) simulations and force fields to characterize structure and dynamics of zinc-binding proteins and investigate the relationship between proteins and metals. Moreover it is important to investigate at molecular level the still unknown effects of cadmium on zinc binding protein, due the relevant implication for human health, since efficient parameters for cadmium are still missing in current force fields. The above considerations prompted us to provide a new approach based on classical molecular mechanics (MM) and quantum chemical calculations, that permits an accurate optimization of force-field parameters for Zn and Cd in CHARMM22/CMAP force field. The new set of parameters was tested investigating the dynamics of SUPERMAN protein zinc-finger domain (SUP_zf) from *A.Thaliana* since two structures were solved by NMR spectroscopy: the wild type one containing zinc (SUP-zf-Zn, PDB entry 1NJQ, Isernia et al., 2003) and one in which zinc is replaced by cadmium, (SUP-zf-Cd, PDB entry 2L1O, Malgieri et al., 2011) (Figure 1) In proteins, Zn can have a catalytic role, by participating directly in catalysis, or a structural role to maintain protein structure and stability. Considering this second role the largest class of Zn-proteins are the Zn-finger and Zn-finger like families, including several transcription factors and the oncosuppressor p53 (Berg, et al., 1987). The zinc-finger domain is a highly conserved domain in which zinc is coordinated by two cysteines and two histidines ($ZnCys_2His_2$) (Hanas et al., 1983). In particular, the classical zinc finger domain, like SUP-zf, is characterized by two β strands, one α helix, and an hairpin structure. During the folding process of zinc finger domains, the Zn atom play a crucial role by driving the folding of the protein and stabilizing the β-hairpin and α-helix. Moreover the metal atom has a crucial structural role in maintaining the integrity of zinc-finger fold and in modulating its dynamics.

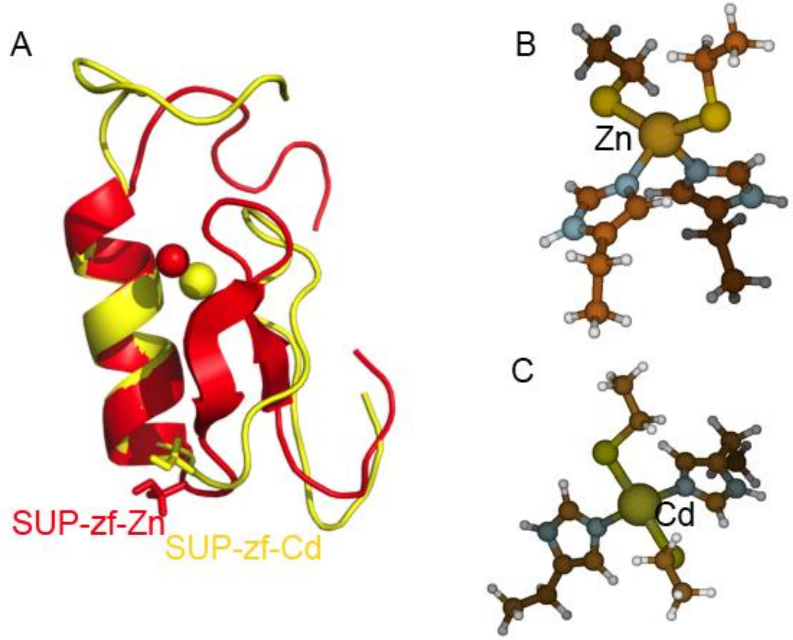


Figure 1. Structure of the SUPERMAN protein zinc finger domain (A). DFT optimized geometries of the [Zn(SEt)$_2$(EtImid)$_2$] (B) and [Cd(SEt)$_2$(EtImid)$_2$] (C)

Ab inito calculations

New parameters for Zn and Cd in the zinc finger domain coordination site were developed using as reference minimum energy geometries of model complexes obtained by quantum mechanics (QM) calculations. QM calculations were performed in collaboration with Prof. Maurizio Bruschi of the Department of Environmental and Earth Sciences at the University of Milano Bicocca. DFT calculations were carried out on the model complexes of the ZnCys$_2$His$_2$ and CdCys$_2$His$_2$ protein binding site [Zn(SEt)$_2$(EtImid)$_2$] and [Cd(SEt)$_2$(EtImid)$_2$], respectively, in which the side chains of Cys and His residues were modelled as ethylthiolate (EtS) and ethylimidazole (EtImid). DFT optimized geometries of the [Zn(SEt)$_2$(EtImid)$_2$] and [Cd(SEt)$_2$(EtImid)$_2$] complexes are shown in Figure 1, while energy and relevant geometry parameters of the two complexes are reported in Table 1.

Table 1 *Ab initio* optimized geometries and energies for tetrahedral zinc and cadmium ligand complexes.

	$[Zn(SEt)_2(EtImid)_2]$	$[Cd(SEt)_2(EtImid)_2]$
Zn(Cd)-S1 (Å)	2.311	2.478
Zn(Cd)-S2(Å)	2.305	2.469
Zn(Cd)-N1(Å)	2.152	2.433
Zn(Cd)-N2(Å)	2.168	2.453
S1-C1(Å)	1.859	1.860
S2-C2(Å)	1.857	1.860
S1-Zn(Cd)-S2 (deg)	138.0	150.7
N1-Zn(Cd)-N2 (deg)	100.7	97.9
Zn(Cd)-S1-C1 (deg)	103.8	105.0
Zn(Cd)-S2-C2 (deg)	104.1	104.3
Zn(Cd)-N1-C3 (deg)	125.0	124.7
Zn(Cd)-N1-C3' (deg)	127.8	128.0
Zn(Cd)-N1-C4 (deg)	123.2	120.7
Zn(Cd)-N1-C4' (deg)	130.0	132.7
Zn(Cd)-S1-C1-C5 (deg)	176.1	177.7
Zn(Cd)-S2-C2-C6 (deg)	172.8	171.5
S1-Zn(Cd)-S2-C2 (deg)	73.3	79.4
S2-Zn(Cd)-S1-C1 (deg)	4.8	5.7
N1-Zn(Cd)-C3-C3' (deg)	4.5	-0.5
N2-Zn(Cd)-C4-C4' (deg)	1.8	5.5

In the former complex the two Zn-S distances are equal to about are 2.31 Å; a value which increase to about 2.47 Å when Zn is replaced by Cd in $[Cd(SEt)_2(EtImid)_2]$. The distances between Zn and δ–N of the two ethylimidazoles also increase after replacement of Zn with Cd, from 2.15 Å and 2.16 Å to 2.43 Å and 2.45 Å, respectively. This trend is clearly due to the larger atomic size of Cd with respect to Zn, as well as to the smaller binding energy of the ligands in the case of Cd (*vide infra*). It is interesting to note that the arrangement of ligands around the metal atom in the two complexes differ significantly from the ideal tetrahedral geometry. In fact, in $[Zn(SEt)_2(EtImid)_2]$ the S-Zn-S angle (138.0 degrees) is much larger than the tetrahedral one, whereas the N-Zn-N angle (100.7 degrees) is smaller. This distortion is still more marked in the [Cd(SEt)$_2$(EtImid)$_2$] complex where S-Zn-S and N-Zn-N angles are equal to 150.7 and 97.9 degrees, respectively. This geometric distortion can be explained by considering the strong repulsion between the two anionic thiolate ligands which significantly affects the S-Zn(Cd)-S angle by increasing the separation of the sulphur atoms, and still confirm the importance of using bonding terms in the force field for a proper description of the structure and dynamics of the zinc-finger domain. The other geometrical parameters are similar in the two complexes. In particular, the Zn(Cd)-S-C angle is equal to about 104 degrees, while the Zn(Cd)-N-C and Zn(Cd)-N-C' angles are equal to about 125 and 130 degrees, respectively. It is important to note that several stable conformations of the two complexes were identified, which essentially depend on the values of the S-Zn(Cd)-S-C and Zn(Cd)-S-C-C dihedral angles. In fact, for each of the two dihedral angles, two set of values are representative for the different conformations. In the case of the S-Zn(Cd)-S-C angle, the values are of about 0 and 75 degrees, while for the Zn(Cd)-S-C-C angle the values are in a range close to 90 and 180 degrees. A detailed analysis of the potential energy surface of the two complexes shows that all stable conformations are almost isoenergetic as the largest energy difference is within 1 kcal/mol. The conformation chosen as reference for the force-field parametrization and discussed above is the lowest energy one. Notably, these results are in agreement with the structural distortion of the metal-binding site of SUP-zf observed by comparing the structures of SUP-zf-Zn and SUP-zf-Cd, and in particular the increment in the metal-ligand distances observed in SUP-zf-Cd with respect to SUP-zf-Zn, with His24 featuring the largest displacement by increasing the distances between the metal atom and the ligands.

The NBO charge calculated for Zn in the [Zn(SEt)$_2$(EtImid)$_2$] complex is equal to +1.44, indicating a significant charge transfer from the ligands to the metal atom. In particular, the net charges of the ligands in the complex are equal to -0.75 for the two thiolates, and close to 0 for the two ethylimidazoles, indicating that the charge transfer is totally due to the anionic ligands. A similar charge distribution is observed in [Cd(SEt)$_2$(EtImid)$_2$] where the NBO charge of the metal atom is +1.36 as due to a slightly larger charge transfer from the two thiolate ligands, which in the complex have a net charge of about -0.70.

Potential energy curves were calculated for the dissociation of thiolate and EtImid ligands from the metal atom, in both [Zn(SEt)$_2$(EtImid)$_2$] and [Cd(SEt)$_2$(EtImid)$_2$] complexes, by performing B3LYP/TZVP single point energy calculations on systems in which the Zn(Cd)-N distance of a EtImid ligand, or the Zn(Cd)-S distance of a thiolate ligand are lengthened or shortened with respect to the optimum geometry value with steps of 0.1 Å. Points at 10.0 and 100.0 Å were also collected to consider the complete dissociation of the ligands. The binding energy profiles calculated as a function of the Zn(Cd)-N and Zn(Cd)-S distances are reported in Figure 2.

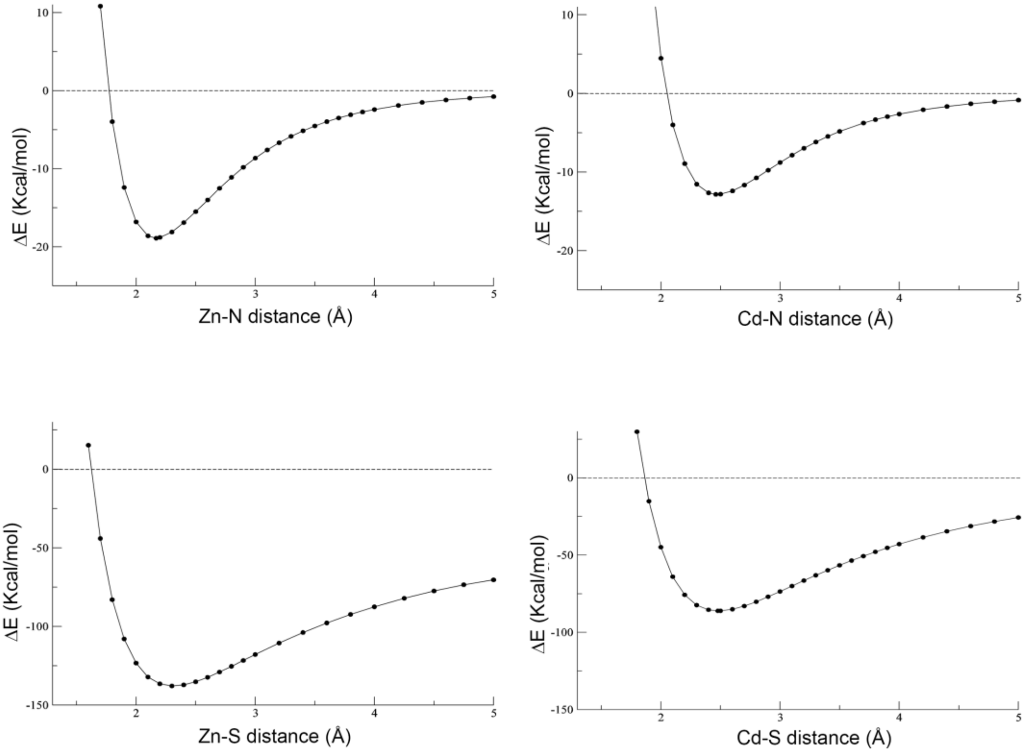


Figure 2. Potential energy curves for the dissociation of metal-ligand in [Zn(SEt)$_2$(EtImid)$_2$] and [Cd(SEt)$_2$(EtImid)$_2$] complexes.

The binding energy of the thiolate ligand in [Zn(SEt)$_2$(EtImid)$_2$] is equal to -137.5 kcal/mol. This value decreases to -84.5 kcal/mol when Zn is replaced by Cd in [Cd(SEt)$_2$(EtImid)$_2$]. The binding energy of the EtImid in [Zn(SEt)$_2$(EtImid)$_2$] is equal to -18.3 kcal/mol, a value much lower than of the thiolate ligand. The binding energy further decreases to -13.1 kcal/mol for the [Cd(SEt)$_2$(EtImid)$_2$] complex. Therefore, the thiolate ligand is bound to metal much more tightly than the imidazole ligands. On the other hand, the energy curve of the Zn(Cd)-thiolate interaction in the dissociation region is much more smooth than that of the Zn(Cd)-imidazole interaction, as illustrated by the fact that the 20% elongation of the Zn-S distance with respect to the optimum value lead to an increase of the energy of less than

10% with respect to the energy minimum, while for the same elongation of the Zn-N distance the increase of the energy is about 35% with respect the energy minimum. This observation is important for the parametrization of the force field as the stretching term is described by an harmonic function and the smoothness of the energy curve is modeled by the value of the force constant.

Molecular mechanics calculations and parameter optimization

The new optimized force field of Zn- and Cd-protein complexes has been derived by using a new approach based on quantum chemical calculations at DFT level, as previously described, and classical molecular mechanics (MM). In our new force field the zinc and cadmium–ligand interactions were modelled with harmonic, bonded terms, in order to consider the covalent character of the metal-ligand bond and prevent the ligands from leaving the coordination sphere.

Table 2. Optimized force-field parameters of Zn- and Cd-protein complexes with the Zn(Cys_2)(His_2) and Cd(Cys_2)(His_2) coordination modes. Labels of atom correspond to the atom typing in CHARMM22/CMAP force field, R_0(Å), θ(deg), K_R K_θ(kcal/mol)

		Zn($His)_2$($Cys)_2$	Cd($His)_2$($Cys)_2$
Stretching term $E = \frac{1}{2}K_R(R - R_0)^2$			
NR2-Zn(Cd)	R_0	2.1683	2.4990
	K_R	150.570	50.010
S-Zn(Cd)	R_0	2.4041	2.5818
	K_R	80.999	88.843
Bending term $E = \frac{1}{2}K_\theta(\theta - \theta_0)^2$			
NR2-Zn(Cd)-NR2	θ	110.5709	104.6806
	K_θ	21.130	62.681
Zn(Cd)-NR2-CPH1	θ	119.8307	114.811
	K_θ	10.4189	14.451
Zn(Cd)-NR2-CPH2	θ	118.4841	10.915
	K_θ	121.979	18.131
NR2-Zn(Cd)-S	θ	114.5330	106.0500
	K_θ	52.817	198.000
S-Zn(Cd)-S	θ	128.165	127.319
	K_θ	103.073	155.291
Zn(Cd)-S-CT2	θ	100.1961	101.0000
	K_θ	60.322	54.940

The new parameters include the stretching terms for the Zn(Cd)-S and Zn(Cd)-N bonds, the bending terms for the S-Zn(Cd)-S, N-Zn(Cd)-N, N-Zn(Cd)-S, Zn(Cd)-S-C and Zn(Cd)-N-C angles, the dihedral terms for the S-Zn(Cd)-S-C and Zn(Cd)-S-C-C angle and the improper dihedral term for the N-Cd(Zn)-C-C angle. The list of the new force-field parameters is reported in Table 2.

The other parameters are those already present in the standard CHARMM22/CMAP force field. In particular, the S-Zn(Cd)-S-C and Zn(Cd)-S-C-C dihedral terms were introduced to explicitly describe the relative orientation between the cysteines in the coordination site, which as stated before span different stable conformations of the model complexes.

In order to obtain optimum values of the force-field parameters we developed a new tool that performs the optimization of the force-field parameters based on the QM results. The program performs iteratively geometry optimization and single point energy calculation with classical molecular mechanics using the program NAMD and the CHARMM22/CMAP force field with the additional terms. The program performs energy calculations on the optimized geometry of the model complexes and on several geometries generated by lengthening or shortening the distances between the metal atom and the S or N atoms of thiolate or imidazole, respectively, in order to reproduce DFT dissociation energy curves. Then, to optimize the force-field parameters and fit QM results the program uses a simulated annealing approach. The simulated annealing is a probabilistic model that permits to locate with good approximation the global minimum of a given error function in a large search space. In the present case the error function to be minimized is given by a weighted sum of the root mean square deviation (rms) between the MM and DFT optimized geometries of the model complexes, and the unsigned sum of difference between the QM and MM interaction energies along the dissociation energy curves in proximity of the energy minimum (five point with shorter distances and step 0.1 Å and five points with longer distances and step 0.1 Å). According to the error value, the simulated annealing approach varies the temperature along each iteration till the system reaches the conformation of minimum energy. Parameters were accepted when the rms deviation error value was under a chosen threshold of 0.01 Å for the geometries and energy error over the dissociation potential energy curves was under 0.1 kcal/mol. The program allowed us to obtain force-field parameters that reproduce DFT geometry and the dissociation curves with good accuracy. Atomic partial charges for the MM calculations were taken directly from the DFT calculations on the $[Zn(SEt)_2(EtImid)_2]$ and $[Cd(SEt)_2(EtImid)_2]$ model complexes. We choose to use NBO charges as this charges are reasonably consistent with the development of the CHARMM force field. In order to take into account polarization and charge transfer effects due to the specific binding site of the metal atom and the nature of the interaction, we developed a program than redistributes residual charges (net charge of the system minus the QM charge of the metal) over a set of heavy atoms in histidine (Nδ, Cδ, Nε, Cε, Cγ) and cysteine (S, Cβ, Cα) by scaling the CHARMM22/CMAP charges of such atoms on the basis of the distance of the atoms from the metal atom. The modified atomic charges of the force field are shown in figure 3. In our new force field the metal-ligand interactions are modeled with harmonic stretching, torsional and bending terms in addition to the classical CHARMM22/CMAP non-bonded terms, in order to consider the covalent character of the metal-ligand bond. Our new approach permits to obtain optimized parameters by accurately reproducing geometry optimized at DFT level and the potential energy curve for the metal-ligands dissociation in the region close to the energy minimum. Firstly we tested the new developed force-field parameters for zinc and Cys_2His_2 coordination sphere and we used them in atomistic explicit-solvent multi-replicate MD simulations using the SUP protein zinc-finger domain (SUP-zf).

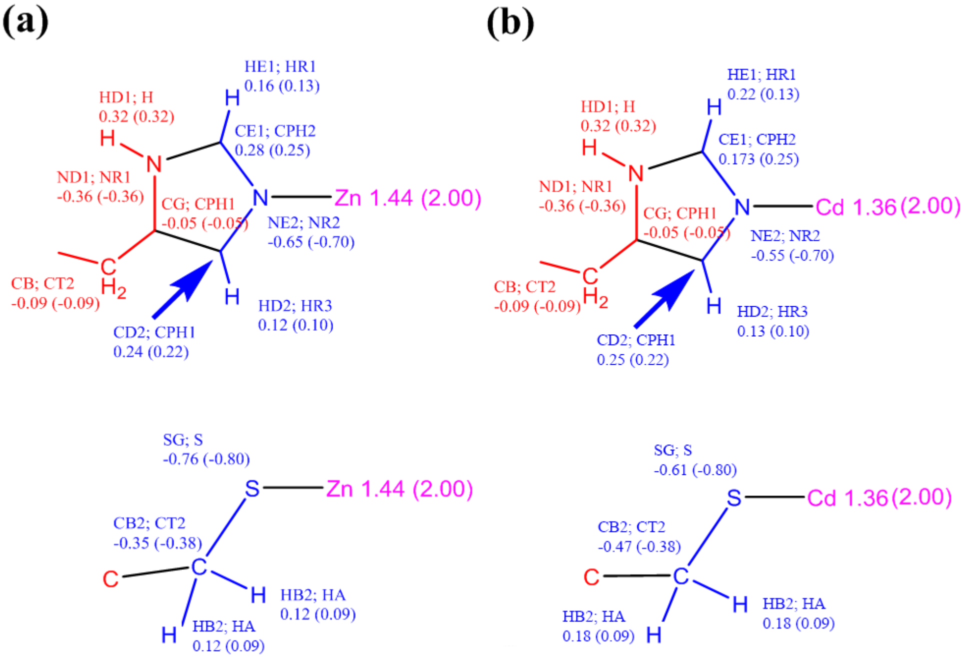


Figure 3. Atomic charges for the force field of (a) Zn- and (b) Cd-protein complexes with the $Zn(Cys_2)(His_2)$ and $Cd(Cys_2)(His_2)$ coordination modes. In parenthesis the charges of the standard CHARMM22/CMAP force field.

Molecular dynamics simulations of Superman protein zinc finger domain

SUP. protein from *Arabidopsis thaliana* is a transcription factor that regulates the expression pattern of homeotic genes in development of flowers. It binds DNA through a single zinc finger domain, in which zinc is coordinated by two cysteines and two histidines. As stated before two structure of SUP-zf have been resolved by NMR techniques: the structure of the native form (SUP-zf-Zn, PDB entry 1NJQ, Isernia et al., 2003) and the structure complexed with Cd (SUP-zf-Cd, PDB entry 2L1O, Malgieri et al., 2011) (Figure 1). NMR structures give information on the structural effects due to zinc to cadmium replacement showing that even if SUP zinc finger domain retains the ββα fold, cadmium induces global structural rearrangements, with changes in the orientation of the secondary structure elements and of the residues essential for DNA recognition and binding (Malgieri et al., 2011). Since the cadmium does not produce global unfold of SUP-zf, it has been proposed that these effects can be reversible and associated with the balance of metals concentration in the cell. The availability of the NMR structure of SUP-zf complexed with zinc and cadmium allowed us to directly check the quality of the new force field, evaluating its ability to yield reliable structures, in good overall agreement with the experimental ones, also when starting from the conformation of SUP-zf containing the different metal atom. In particular, we carried out MD simulations of SUP-zf starting from the native structure (SUP-zf-Zn) and from the Cd-contained structure (SUP-zf-Cd) to test the ability of the new parameters to maintain the observed conformations. In addition, we performed MD simulations of SUP-zf starting from the Cd-contained NMR structure (SUP-zf-Cd), and replacing Cd with Zn (SUP-zf-Cd(Zn)) to check if the parameters can restore conformations close to the native structure. MD simulations were carried out using both the Zn-protein parameters already included in the standard CHARMM22/CMAP force field (in the following such MD simulations will be labelled as SUP-zf-Zn-C22, SUP-zf-Cd(Zn)-C22), and using our new derived Zn-protein parameters (in the following such MD simulations will be labelled as SUP-zf-Zn-new, SUP-zf-Cd(Zn)-new). The results of MD simulations with the two set of parameters have then been compared to investigate their performances in the sampling of conformational ensembles. The evolution of the Cα root mean square deviation (rmsd) over the simulation time, using as reference the NMR structure with zinc (SUP-zf-Zn) is reported in Figure 4. The Cα rmsd required up to 5 ns, which were thus discarded from further analyses, to reach stable values. Rmsd analysis of the MD simulations started from the native structure (SUP-zf-Zn) points out that the NMR resolved structure of SUP-zf is very stable during simulation time, both by using the standard force field (SUP-zf-Zn-C22) and our new derived force field (SUP-zf-Zn-C22-new). Moreover with both force fields, the tetrahedral arrangement of ligands in the coordination site is preserved. This observation indicates that standard force field is sufficient to preserve the structural features of the protein by sampling conformations close to the native structure. However simulations performed with the standard force field, started from the NMR structure of SUP-zf-Cd in which Cd was replaced by Zn (SUP-zf-Cd(Zn)-C22), do not feature the conformational transition to the native structure during the simulation time, as shown by the rmsd values always larger than 0.7 nm.

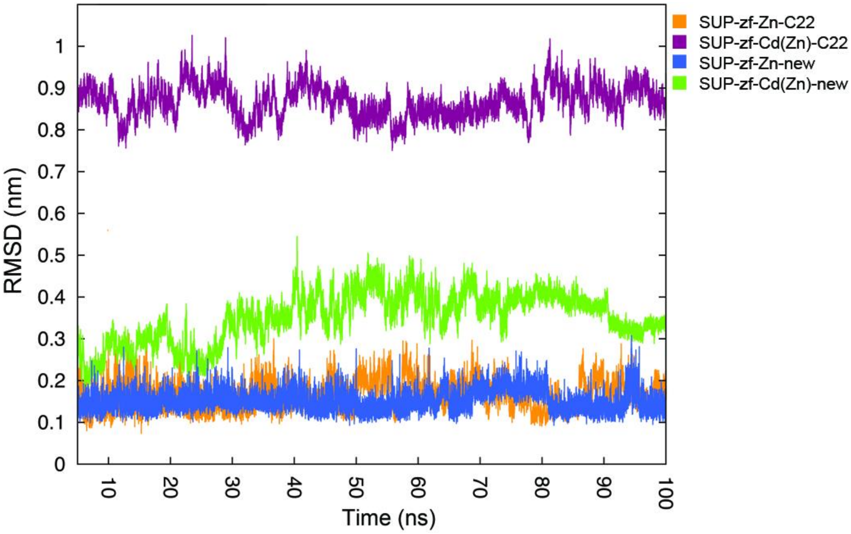


Figure 4. Rms deviations from the experimental structure SUP-zf-Zn for Cα atoms during MD simulations. The overall deviations from the experimental structure are reported as a function of time for each simulation.

Moreover in the simulations performed with the standard force field, an expansion of the coordination sphere of zinc is observed as the metal atom gives an additional interaction with one water molecule, forming a penta-coordinated site, which is not observed in the experimental structure. On the other hand, in comparison with the standard force field, our optimized force field improve the agreement of the MD sampled conformations with the NMR-resolved structure, showing very small overall deviations (average rmsd 0.15 ± 0.07 nm). In addition, when starting MD simulations from the NMR structure of SUP-zf-Cd and replacing Cd with Zn (SUP-zf-Cd(Zn)-C22-new), the conformational transition to the native structure occurs during the simulation time, as the average rmsd is 0.25 ± 0.04 nm. Furthermore, with the our optimized force field the tetrahedral geometry of Zn is preserved during all the simulation time, without any expansion of the coordination sphere. To better describe the efficacy of the new derived force field to model the conformational changes induced after the insertion of Zn in SUP-zf-Cd, we examined the evolution in the values of selected angles that can better describe conformational changes caused by cadmium in SUP-zf (Figure 5). Indeed, the angles defined by the Cα atoms of residues Gly63, His65 and Lys53 and the Cα atoms of residues Gly63, His65 and Cys49 can accurately describe the global structural perturbations in the secondary structure of the protein, and in the metal coordination sphere induced by cadmium insertion in SUP-zf-Cd with respect to the native structure. These angles are therefore used to discriminate between SUP-zf-Zn and SUP-zf-Cd structures. In fact, in the NMR structure of the native protein (SUP-zf-Zn) the Gly63-His65-Lys53 and Gly63-His65-Cys49 angles are equal to 116.6 and 95.2 degrees, respectively, while in the NMR structure of the Cd-containing protein (SUP-zf-Cd) they are 92.5 and 73.2 degrees, respectively. This values show that a rotation of more than 20 degrees in the orientation of secondary structure elements around the metal coordination site occurs in the structures of SUP-zf-Cd compared to that of SUP-zf-Zn (Figure 5). In the simulations performed starting from the native structure, and using both the standard and the new derived force fields (SUP-zf-Zn-C22 and SUP-zf-Zn-new, respectively), variation of these angles is very large, showing that the structure of SUP-zf has a certain degree of flexibility. In addition, variations of the angles are similar in the MD simulations performed with the two force fields, suggesting that with both force fields SUP-zf can maintain the structural integrity of the protein. On the contrary, in SUP-zf-Cd(Zn)-C22 simulations (with the standard CHARMM/CMAP force field) the average Gly63-His65-Lys53 and Gly63-His65-Cys49 angles are 107.6 ± 4.9 and 94.5 ± 5.04 degrees, confirming that conformations approaching the experimental SUP-zf-Zn structure are not sampled during simulation, demonstrating the limit in the correct description of the conformational modifications in the protein. Notably, the introduction of our optimized parameters in SUP-zf-Cd(Zn)-new simulations permits to sample conformations very close to the experimental structure with average Gly63-His65-Lys53 and Gly63-His65-Cys49 angles of 117.6 ± 5.23 ° and 95.7 ± 5.98 ° (Figure 5) significantly improving the MD ensemble and showing the reconstitution of native-like structures. In summary, our new force field allows, in MD simulations, to reproduce the conformational changes in SUP-zf induced by the substitution of Cd with Zn and observed with NMR spectroscopy, in particular, by restoring the orientation of secondary structure elements.

The NMR structure shows that cadmium induces also perturbations in the orientation of Ser58 side chain, which has been proposed to play a key role in the DNA recognition. The side chain of this residue, in fact, is outward facing in the SUP-zf-Zn while is oriented towards in the SUP-zf-Cd. We calculate the rmsd of the side chain atoms of Ser58 relatively to native state to evaluate its orientation during SUP-zf-Cd(Zn)-c22 and SUP-zf-Cd(Zn)-new simulations (Figure 5). The analysis shows that only with our new parameters Ser58 return to a functional orientation similar to the SUP-zf-Zn (average rmsd 0.17 ± 0.03 nm), facing outward the domain and exposed to interact with DNA.

In order to further check if SUP-zf-Cd(Zn)-new simulations sample conformations close to that of the NMR SUP-zf-Zn structure, the different classes of intramolecular interactions were calculated in the MD *ensemble* along with their persistence (Figure 6).

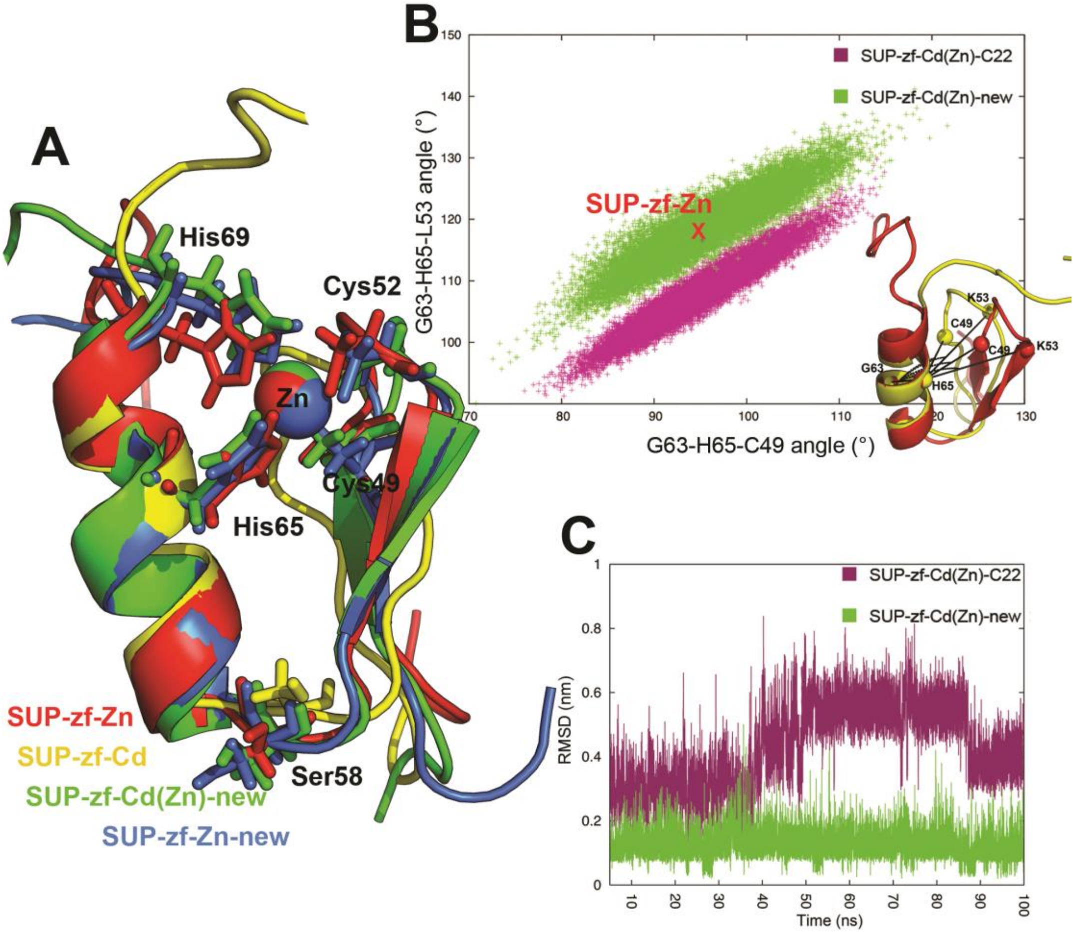


Figure 5. Reproduction of experimentally observed conformational changes during MD simulations. Only our novel parameters permit to restore the native structure and conformational changes of SUP-zf observed by NMR spectroscopy during MD simulations conducted starting from the cadmium-distorted structure. A) Structural comparison between structures sampled during MD simulations after 100 ns of simulation time (SUP-zf-Zn-new and SUP-zf-Cd(Zn)-new) and the NMR structures (SUP-zf-Zn and ZUP-zf-Cd). B) evolution of the Gly63-His65-Lys53 and Gly63-His65-Cys49 angles during MD simulations. The red mark indicates the values of these angles in the native structure (SUP-zf-Zn). C) Rmsd profile for the side chain atoms of the Ser58 using as reference the is position in the native structure (SUP-zf-Zn).

The networks of electrostatic, hydrogen bonds and hydrophobic interactions were calculated using *PyInteraph* (Tiberti et al., 2014) and analyzed with Xpyder tools (Pasi et al., 2012). In SUP-zf the hydrophobic core composed by Phe56, Leu62, Ala61 and Tyr27 plays a crucial role in maintaining thermal stability and 3D structure (Malgieri et al., 2011). Moreover zinc-finger domains have a typical pattern of hydrogen bonds associated with the conservation of the fold and the secondary structures. Our analysis shows highly persistent interactions among Phe56, Leu62 and Ala61 of the hydrophobic core, in agreement with experimental data, which should be crucial in maintaining the global structural integrity of the domain during MD simulations. Moreover a wide network of hydrogen bonds is present and several such bonds are highly persistent during SUP-zf-Cd(Zn)-new simulations. In particular, the analysis shows that selected hydrogen bond associated with the maintenance of the α helix (among residues from Ala61 to Met66), and the two β-hairpin (among residues from Cys49 to Arg54), are highly persistent in our MD ensemble (more than 70 %). Interestingly, this result is in agreement with the network of hydrogen bonds observed in most of the conformers of the deposited NMR structures.

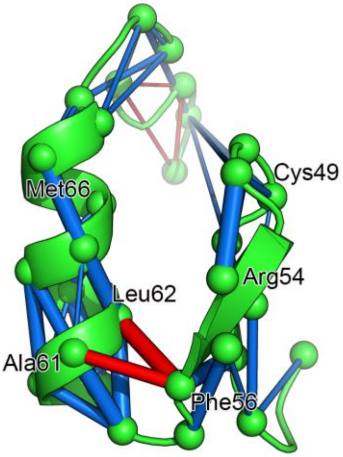


Figure 6. Hydrophobic interactions and hydrogen bonds involved in the stability of SUP-zf core and secondary structures. We used a significant threshold of persistence of 20%, to discard low populated interactions in the ensemble. The SUP-zf structure is shown in green cartoon. The Cα atoms of the residues involved in the interactions are indicated as spheres, respectively. The hydrophobic interactions are represented as red cylinders while hydrogen bonds as blue cylinders, connecting the Cα atoms of residues and their thickness is proportional to the persistence value.

The very good agreement with the experimental data of the MD simulations performed with our new parametrized force field support the efficacy of our approach. It remains to investigate whether repeated and longer simulations of these proteins with the new force field would consistently give the same results. We are now collecting microsecond MD simulations to further investigate properties of protein ensembles (Dror et al., 2012) and metadynamics that will allow us to describe the free energy landscape associated with conformational changes caused by metal replacement (Sutto et al., 2012).

Molecular Dynamics simulations of SUP-zf in the presence of cadmium with optimized parameters

Using the fitting approach discussed in the previous section new derived force field parameters for the interaction of cadmium and Cys_2His_2 coordination geometry and we tested them in atomistic explicit-solvent MD simulation of 100 ns. We performed MD simulations using the SUP-zf-Zn structure solved by NMR in which we manually replaced the zinc with cadmium (SUP-zf-Zn(Cd)) to evaluate if with our optimized parameters they can describe and resemble conformational alterations induced by cadmium and observed experimentally. To compare our results since no parameters for cadmium are present in CHARMM22/CMAP we used a set of parameters previously obtained to reproduce QM calculations of cadmium in water (Li et al. 2014). To evaluate the effectiveness of our new parameters to sample altered structures after the substitution with cadmium, resembling those observed with NMR spectroscopy, we monitored in the MD ensemble the angles previously described (Figure 7). The angles between Cα of residues Gly63, His65 and Lys53 and of residues Gly63, His65 and Cys49 in the NMR ensemble of the SUP-zf-Cd measure 92.5 ± 2.3 and 73.2 ± 1.4 degrees, respectively (Figure 7). In SUP-zf-Zn(Cd)-c22 simulations with the Charmm22/CMAP and old parameters the analysis of the 2D subspace described by the two angles points out that at least two substates of SUP-zf can be observed: one, most populated, has average Gly63-His65-Lys53 and Gly63-His65-Cys49 angles of 135.7 ± 7.9 and 117.5 ± 8.4 degrees and one, less populated, has average angles 109.5 ± 4.9 and 92.5 ± 4.1 degrees. These angles are very different from those measured in the NMR structure of SUP-zf-Cd showing that states resembling the experimental SUP-zf-Cd structure are not sampled during MD simulation. These results point out that the old set of parameters is not able to describe dynamics and conformational alteration induced by cadmium in the protein reliably. On the contrary, the introduction of our optimized force field parameters in SUP-zf-Zn(Cd)-new simulations permits to sample conformations more close to the experimental structure. In fact in the 2D subspace described by the two angles only one populations can be seen with average Gly63-His65-Lys53 and Gly63-His65-Cys49 angles of 102.5 ± 5.7 and 86.7 ± 4.6 degrees (Figure 7) clearly improving the agreement with the cadmium altered structure of SUP-zf-Cd.

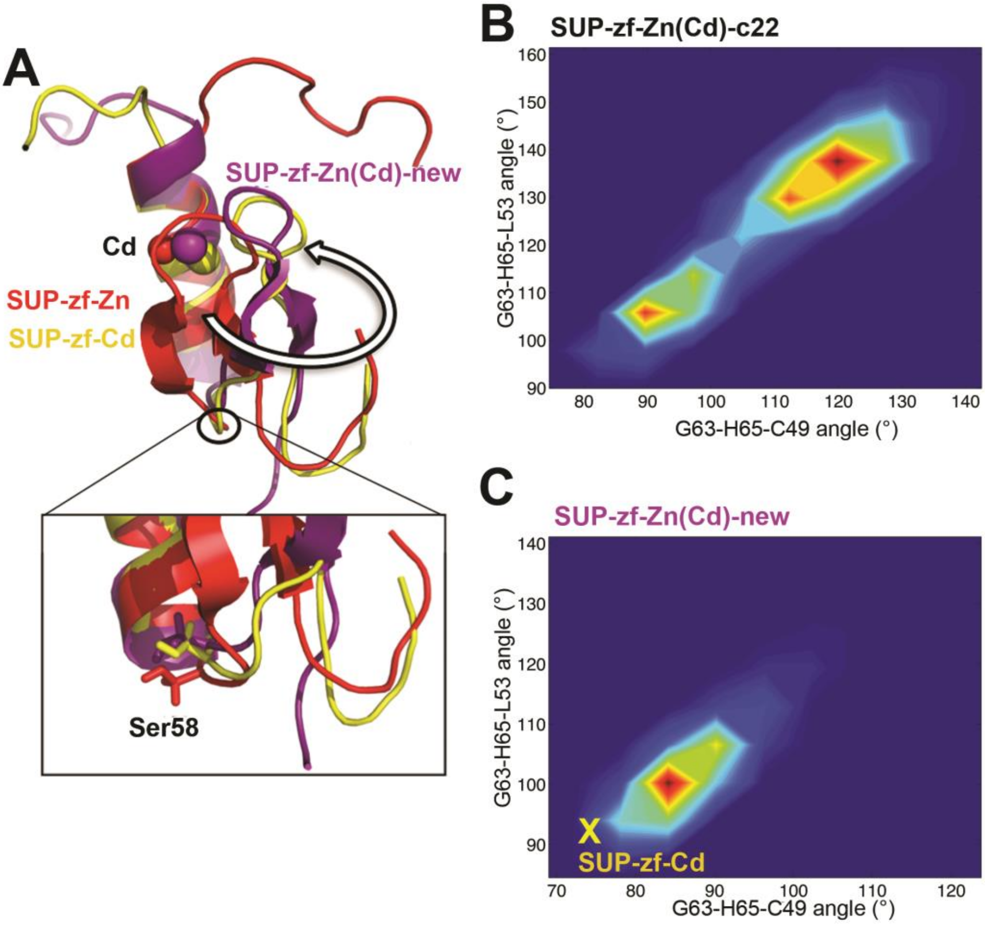


Figure 7. Reproduction of experimentally observed conformational changes during MD simulations. Only our novel parameters permit to sample conformations similar to the structure of SUP-zf-Cd during MD simulations conducted starting from the native structure. A) Structural comparison between structures sampled during MD simulations after 100 ns of simulation time (SUP-zf-Zn(Cd)-new) and the NMR structures (SUP-zf-Zn and ZUP-zf-Cd). B and C show the values for the Gly63-His65-Lys53 and Gly63-His65-Cys49 angles during MD simulations with old (SUP-zf-ZN(Cd)) and our new parameters (SUP-zf-Zn(Cd)-new). The two-dimensional probability density distribution of the two angles is here shown as a contour plot from blue (low populated regions) to red (highly populated regions).

In addition, MD simulations performed with the new derived force field parameters reproduce conformational changes in zinc-finger domain as those induced by the substitution of Zn with Cd observed with NMR spectroscopy, in particular, by restoring orientation of secondary structure elements. These preliminary results support the potential of our approach for the development of force field parameters to model metal binding sites in proteins. However, improvement in the definition of these parameters is still necessary since structures that completely resemble the SUP-zf-Cd are not sampled.

3.2.4 Conclusions

Binding of zinc in proteins is essential since it serves multiple essential biological processes but the exact mechanism and physicochemical principles governing protein-metal relationship remain elusive (Dudev et al., 2008, Andreini et al., 2006). Moreover alteration of zinc homeostasis in cells by cadmium causes several pathological states, like cancer (Loh et al., 2010, Breydo et al., 2011, Xu et al., 2011) but little is known about the cadmium interferences with metal binding and how it induces development of diseases. Due to the complexity of the problem robust experimental analysis are missing suggesting that an *in silico* approach is essential in order to investigate dynamics and structural properties of metal binding proteins. Several approaches associated with Molecular Dynamics simulations have been proposed but they still have approximations to characterize zinc-binding proteins and investigate the relationship between proteins and metals (Calimet et al., 2006, Hu et al., 2011). The above considerations prompted us to develop a new tool based on classical molecular mechanics (MM) and quantum chemical calculations at the density functional theory (DFT) level, that permits to develop optimized parameters for the metal ions, starting from the parameters from CHARMM22/CMAP force field, that can be readily used in MD simulations to describe the coordination of metals in metal-binding proteins. In our new force field we introduced new terms to accurately describe the metal coordination site and take into account the covalent character of the metal-ligand bond and polarization effects due to different coordination elements. We preliminary examined the behavior of our optimized parameters in atomistic explicit-solvent MD simulations on the zinc finger domain of SUPERMAN protein and compared our results with structures previously obtained by NMR techniques pointing out the relevancy of our new approach. We shown that actual CHARMM22/CMAP that has only nonbonded terms to treat the coordination with the metal ions in proteins, leads to poor accuracy for the association energies and despite it permits to retain the stability and ββα fold of SUP-zf, it leads to structures where the tetrahedral metal coordination is lost. The use of additional terms and of our forcefield parameters optimized to reproduce QM calculations, permits to sample structures that are in good agreement with the structures experimentally observed. Only with our parameters MD simulations starting from the NMR structure of SUP-zf solved in the presence of cadmium in which the zinc was reintroduced succeed in reconstitute the native structure. Our approach is simple but straightforward and our preliminary results suggest that it is an useful tool to study through MD simulations metal binding proteins in solution. Our approach is not suitable to study effects that involve the binding or unbinding with metals ions like in the unfolding/folding of proteins or when the zinc coordination is not maintained. In order to describe such processes, more complex approaches have to be taken into account, like the use of polarizable force fields. Our preliminary investigation strongly support the reliability of our approach to study effects induced by metal ions on overall proteins dynamics and structures, demonstrating that also fine and very specific conformational alterations can be very well described, as for example the structural modifications induced by cadmium. Moreover we are further testing the consistency of our approach, collecting microsecond MD simulations to further investigate properties of protein ensembles (Dror et al., 2012) and metadynamics that will permit to describe the free energy landscape associated with conformational changes caused by metal replacement (Sutto et al. 2012). Our computational approach will have an important impact, opening new opportunity to investigate dynamics and functional properties of other zinc-binding proteins. Moreover the application of our approach to develop new parameters for pollutant metals will permit to provide insights into the molecular basis of the development of disease.

3.2.5 References

Andreini, C., L. Banci, I. Bertini and A. Rosato (2006). Counting the zinc-proteins encoded in the human genome. *J Proteome Res* 5(1): 196-201.

Becke, A.D. (1993) Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* 98, 5648 - 5652.

Banci, L. (2003). Molecular dynamics simulations of metalloproteins. *Curr Opin Chem Biol* 7(1): 143-149.

Berg, J. M. and H. A. Godwin (1997). Lessons from zinc-binding peptides. *Annu Rev Biophys Biomol Struct* 26: 357-371.

Breydo, L. and V. N. Uversky (2011). Role of metal ions in aggregation of intrinsically disordered proteins in neurodegenerative diseases. *Metallomics* 3(11): 1163-1180.

Buchko, G. W., N. J. Hess and M. A. Kennedy (2000). Cadmium mutagenicity and human nucleotide excision repair protein XPA: CD, EXAFS and (1)H/(15)N-NMR spectroscopic studies on the zinc(II)- and cadmium(II)-associated minimal DNA-binding domain (M98-F219). *Carcinogenesis* 21(5): 1051-1057.

Calimet, N. and T. Simonson (2006). Cys(x)His(y)-Zn2+ interactions: possibilities and limitations of a simple pairwise force field. *J Mol Graph Model* 24(5): 404-411.

Chakravorty, D. K., B. Wang, C. W. Lee, D. P. Giedroc and K. M. Merz, Jr. (2012). Simulations of allosteric motions in the zinc sensor CzrA. *J Am Chem Soc* 134(7): 3367-3376.

Dal Peraro, M., K. Spiegel, G. Lamoureux, M. De Vivo, W. F. DeGrado and M. L. Klein (2007). Modeling the charge distribution at metal sites in proteins for molecular dynamics simulations. *J Struct Biol* 157(3): 444-453.

Donini, O. A. and P. A. Kollman (2000). Calculation and prediction of binding free energies for the matrix metalloproteinases. *J Med Chem* 43(22): 4180-4188.

Dror, R. O., R. M. Dirks, J. P. Grossman, H. Xu and D. E. Shaw (2012). Biomolecular simulation: a computational microscope for molecular biology. *Annu Rev Biophys* 41: 429-452.

Dudev, T. and C. Lim (2008). Metal binding affinity and selectivity in metalloproteins: insights from computational studies. *Annu Rev Biophys* 37: 97-116.

Elstner, M., Q. Cui, P. Munih, E. Kaxiras, T. Frauenheim and M. Karplus (2003). Modeling zinc in biomolecules with the self consistent charge-density functional tight binding (SCC-DFTB) method: applications to structural and energetic analysis. *J Comput Chem* 24(5): 565-581.

Hanas, J. S. and C. G. Gunn (1996). Inhibition of transcription factor IIIA-DNA interactions by xenobiotic metal ions. *Nucleic Acids Res* 24(5): 924-930.

Hanas, J. S., D. J. Hazuda, D. F. Bogenhagen, F. Y. Wu and C. W. Wu (1983). Xenopus transcription factor A requires zinc for binding to the 5 S RNA gene. *J Biol Chem* 258(23): 14120-14125.

Hartwig, A. (2001). Zinc finger proteins as potential targets for toxic metal ions: differential effects on structure and function. *Antioxid Redox Signal* 3(4): 625-634.

Hartwig, A. (2010). Mechanisms in cadmium-induced carcinogenicity: recent insights. *Biometals* 23(5): 951-960.

Hartwig, A., M. Asmuss, H. Blessing, S. Hoffmann, G. Jahnke, S. Khandelwal, A. Pelzer and A. Burkle (2002). Interference by toxic metal ions with zinc-dependent proteins involved in maintaining genomic stability. *Food Chem Toxicol* 40(8): 1179-1184.

Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. (1997). LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* 18, 1463-1472.

Hu L., U. Ryde, (2011), Comparison of Methods to Obtain Force-Field Parameters for Metal Sites. Journal of Chemical Theory and Computation 7, 2452-2463.

Huang, M., D. Krepkiy, W. Hu and D. H. Petering (2004). Zn-, Cd-, and Pb-transcription factor IIIA: properties, DNA binding, and comparison with TFIIIA-finger 3 metal complexes. *J Inorg Biochem* 98(5): 775-785.

Isernia, C., E. Bucci, M. Leone, L. Zaccaro, P. Di Lello, G. Digilio, S. Esposito, M. Saviano, B. Di Blasio, C. Pedone, P. V. Pedone and R. Fattorusso (2003). NMR structure of the single QALGGH zinc finger domain from the Arabidopsis thaliana SUPERMAN protein. *Chembiochem* 4(2-3): 171-180.

Klepeis, J. L., K. Lindorff-Larsen, R. O. Dror and D. E. Shaw (2009). Long-timescale molecular dynamics simulations of protein structure and function. *Curr Opin Struct Biol* 19(2): 120-127.

Kopera, E., T. Schwerdtle, A. Hartwig and W. Bal (2004). Co(II) and Cd(II) substitute for Zn(II) in the zinc finger derived from the DNA repair protein XPA, demonstrating a variety of potential mechanisms of toxicity. *Chem Res Toxicol* 17(11): 1452-1458.

Kuppuraj, G., M. Dudev and C. Lim (2009). Factors governing metal-ligand distances and coordination geometries of metal complexes. *J Phys Chem B* 113(9): 2952-2960.

Kuwahara, J. and J. E. Coleman (1990). Role of the zinc(II) ions in the structure of the three-finger DNA binding domain of the Sp1 transcription factor. *Biochemistry* 29(37): 8627-8631.

Lee, C., W. Yang, R.G., Parr,Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density *Phys. Rev. B*, 1988, *37*, 785-789

Li, P., and K. M., Merz, (2014). Taking into Account the Ion-Induced Dipole Interaction in the Nonbonded Model of Ions. *J. Chem. Theory Comput.*, 10 (1):289–297

Li, Y. L., Y. Mei, W. Zhang da, D. Q. Xie and J. Z. Zhang (2011). Structure and dynamics of a dizinc metalloprotein: effect of charge transfer and polarization. *J Phys Chem B* 115(33): 10154-10162.

Lindorff-Larsen, K., N. Trbovic, P. Maragakis, S. Piana and D. E. Shaw (2012). Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *J Am Chem Soc* 134(8): 3787-3791.

Loh, S. N. (2010). The missing zinc: p53 misfolding and cancer. *Metallomics* 2(7): 442-449.

Mackerell, A. D., Jr., M. Feig and C. L. Brooks, 3rd (2004). Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem* 25(11): 1400-1415.

Makowski, G. S. and F. W. Sunderman, Jr. (1992). The interactions of zinc, nickel, and cadmium with Xenopus transcription factor IIIA, assessed by equilibrium dialysis. *J Inorg Biochem* 48(2): 107-119.

Malgieri, G., L. Zaccaro, M. Leone, E. Bucci, S. Esposito, I. Baglivo, A. Del Gatto, L. Russo, R. Scandurra, P. V. Pedone, R. Fattorusso and C. Isernia (2011). Zinc to cadmium replacement in the A. thaliana SUPERMAN Cys(2) His(2) zinc finger induces structural rearrangements of typical DNA base determinant positions. *Biopolymers* 95(11): 801-810.

Martelli, A., E. Rousselet, C. Dycke, A. Bouron and J. M. Moulis (2006). Cadmium toxicity in animal cells by interference with essential metals. *Biochimie* 88(11): 1807-1814.

Miller, Y., B. Ma and R. Nussinov (2010). Zinc ions promote Alzheimer Abeta aggregation via population shift of polymorphic states. *Proc Natl Acad Sci U S A* 107(21): 9490-9495.

Parr, R. G., W., Yang.(1989). Density Functional Theory of Atoms and Molecules; Oxford University Press: New York.

Pang, Y. P. (2001). Successful molecular dynamics simulation of two zinc complexes bridged by a hydroxide in phosphotriesterase using the cationic dummy atom method. *Proteins* 45(3): 183-189.

Papaleo, E., K. Lindorff-Larsen and L. De Gioia (2012). Paths of long-range communication in the E2 enzymes of family 3: a molecular dynamics investigation. *Phys Chem Chem Phys* 14(36): 12515-12525.

Reed, A. E., L. A., Curtiss, and F. Weinhold, Intermolecular Interactions from a Natural Bond Orbital, Donor-Acceptor Viewpoint, *Chem. Rev.* 88, 899-926 (1988);

Roesijadi, G., R. Bogumil, M. Vasak and J. H. Kagi (1998). Modulation of DNA binding of a tramtrack zinc finger peptide by the metallothionein-thionein conjugate pair. *J Biol Chem* 273(28): 17425-17432.

Russell, A. J. and A. R. Fersht (1987). Rational modification of enzyme catalysis by engineering surface charge. *Nature* 328(6130): 496-500.

Ryde, U. (1995). Molecular dynamics simulations of alcohol dehydrogenase with a four- or five-coordinate catalytic zinc ion. *Proteins* 21(1): 40-56.

Sakharov, D. V. and C. Lim (2009). Force fields including charge transfer and local polarization effects: Application to proteins containing multi/heavy metal ions. *J Comput Chem* 30(2): 191-202.

Seneque, O. and J. M. Latour (2010). Coordination properties of zinc finger peptides revisited: ligand competition studies reveal higher affinities for zinc and cobalt. *J Am Chem Soc* 132(50): 17760-17774.

Schäfer A., H., Horn, R., Ahlrichs (1992) Fully optimized contracted Gaussian basis sets for atoms Li to Kr. *J Chem Phys* 97(4):2571–2577

Schäfer A., C., Huber, R., Ahlrichs (1994) Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr. *J Chem Phys* 100(8):5829–5835

Stote, R. H. and M. Karplus (1995). Zinc binding in proteins and solution: a simple but accurate nonbonded representation. *Proteins* 23(1): 12-31.

Sutto, L., M. Suppo, F.L. Gervasio (2012). New advances in metadynamics. *Wiley Interdiscip Rev Comput Mol Sci*(2): 771-779.

Tang, J., S. G. Kang, J. G. Saven and F. Gai (2009). Characterization of the cofactor-induced folding mechanism of a zinc-binding peptide using computationally designed mutants. *J Mol Biol* 389(1): 90-102.

Thevenod, F. (2010). Catch me if you can! Novel aspects of cadmium transport in mammalian cells. *Biometals* 23(5): 857-875.

Tiberti, M., G. Invernizzi, M. Lambrughi, Y. Inbar, G. Schreiber and E. Papaleo (2014). PyInteraph: a framework for the analysis of interaction networks in structural ensembles of proteins. *J Chem Inf Model* 54(5): 1537-1551.

Toba, S., K. V. Damodaran and K. M. Merz, Jr. (1999). Binding preferences of hydroxamate inhibitors of the matrix metalloproteinase human fibroblast collagenase. *J Med Chem* 42(7): 1225-1234.

Waisberg, M., P. Joseph, B. Hale and D. Beyersmann (2003). Molecular and cellular mechanisms of cadmium carcinogenesis. *Toxicology* 192(2-3): 95-117.

Wise-Scira, O., L. Xu, G. Perry and O. Coskuner (2012). Structures and free energy landscapes of aqueous zinc(II)-bound amyloid-beta(1-40) and zinc(II)-bound amyloid-beta(1-42) with dynamics. *J Biol Inorg Chem* 17(6): 927-938.

Xu, J., J. Reumers, J. R. Couceiro, F. De Smet, R. Gallardo, S. Rudyak, A. Cornelis, J. Rozenski, A. Zwolinska, J. C. Marine, D. Lambrechts, Y. A. Suh, F. Rousseau and J. Schymkowitz (2011). Gain of function of mutant p53 by coaggregation with multiple tumor suppressors. *Nat Chem Biol* 7(5): 285-295.

Zhang, J., W. Yang, J. P. Piquemal and P. Ren (2012). Modeling Structural Coordination and Ligand Binding in Zinc Proteins with a Polarizable Potential. *J Chem Theory Comput* 8(4): 1314-1324.

Zhang, W., T. J. Hou, X. B. Qiao, S. Huai and X. J. Xu (2004). Binding affinity of hydroxamate inhibitors of matrix metalloproteinase-2. *J Mol Model* 10(2): 112-120.

4 Concluding remarks

The projects discussed in the present thesis show that molecular dynamics simulations performed with state-of-the-art force fields, if integrated with experimental biophysical techniques, are a powerful and effective tool to describe structural and dynamic properties of even complex protein systems, such as intrinsically disordered or metal-binding domains. Moreover, our study points out that the use of methods inspired by graph theory to analyse results from molecular simulations are an effective approach, to investigate structural communication and to predict effects induced over long distances in the protein upon interactions with other biological partners, such as for example DNA, or metal cofactors.

Our results open new questions on the systems that we have been studied, questions that can be addressed by future computational and experimental investigations. For example, we here predict that the regions in the surrounding of loop S6-S7 of p53 DNA Binding domain (DBD) can be an interface to recruit other biological partners involved in p53 signalling and we identified the domains of the aforementioned partners, which are expected to be involved in the interaction. We thus provide the community with a dataset of protein domains for which the binding with p53 can be experimentally validated for example by chemical shift perturbation experiments with NMR or using experimental mutagenesis of residues at the predicted interaction interface and cross-linking experiments. Moreover, the low populated intramolecular interactions identified in intrinsically disordered (ID) domains, such as Sic1 KID domain or AT3$_{182-291}$ fragment, represent suitable candidates for experimental mutagenesis aimed at further characterizing their role in the structural properties of these domains.

Furthermore, our results also highlight limitations that are still present in the classical force fields for MD and suggest new directions for future improvements. Indeed, we here show that a tight integration of MD simulations with data from different experimental biophysical techniques can help in overcoming these limitations. In fact it is known that the accuracy of simulations is critically depending on the physical model (force field) used to describe the simulated molecular system and they are still far from being perfect and not always enough accurate for the case of study. For example, it has recently showed that classical MD force fields overestimate compact states in IDPs, and our results on the disordered regions of AT3 confirm this scenario. Indeed, we show that even state-of-the-art force field for folded proteins are limited in the study of IDPs by the risk to populate unrealistic over-collapsed states, as the ones observed for a minor population in the simulated ensemble of Sic1 kinase inhibitory domain and the C-terminal domains of Ataxin 3. Moreover, even force fields that do not overestimate collapsed states might suffer of other problems when it comes to the study of IDPs, as we here show for the C-terminal ubiquitin interacting motif of Ataxin-3. Indeed, one of the two force fields employed provide reasonably expanded states for the fragment but overestimate the helical population. Overall, MD results if not accurately compared with experimental data risk to provide erroneous conclusions. We compared and integrated our simulations with data from different experimental biophysical techniques to overcome the limitations inherent in classical MD. In particular, NMR spectroscopy is a suitable technique to integrate and validate the results from MD simulations since it has the potential to provide atomistic information on highly dynamic systems and to identify both short range and long-range effects, as we here showed in our study of the ID domain of AT3

Moreover, we know that canonical MD simulations, also when carried out for milliseconds encounter the risk to be entrapped in local minima without allowing to sample the whole conformational space, providing a limited description of the dynamics of the protein. To overcome intrinsic limitations in the conformational sampling of classical MD, we employed enhanced sampling techniques integrated to a state-of-the-art force field that permitted us to describe the conformational changes of the p53 DNA binding domain after the interaction with DNA and obtain quantitative information of the effects induced on its free energy landscape.

Our results also show that approximations are present in the current force fields when the coordination of a protein with metal ions is described. Classical MD force fields, indeed, do not explicitly take into account polarizable effects that are necessary to describe with accuracy metal-protein interactions. To overcome this problem, we developed a new tool based on classical molecular mechanics and quantum chemical calculations that permits to develop optimized parameters for the metal ions, which can be then used in MD simulations. Our optimized parameters show a better agreement with experimental data than the available parameters for metal ions but have to be further tested and validated in the next future.

In conclusion, our study provide examples for future directions in the field of protein structural biology toward the goal of integrating simulations and experimental biophysical techniques to address complicate cases of study, such as metallo-proteins and intrinsically disordered domains. The combination of these approaches is certainly the better valuable mean to achieve and accurate and atomic-level description of protein dynamics.

. .

4.1 Attachments

The following articles will be sent as distinct files:

Intramolecular interactions stabilizing compact conformations of the intrinsically disordered kinase-inhibitor domain of Sic1: a molecular dynamics investigation.
Lambrughi M., Papaleo E., Testa L., Brocca S., De Gioia L. and Grandori R. (2012)
Front. Physiol., 3:435.

The conformational ensemble of the disordered and aggregation-protective 182-291 region of ataxin-3.
Invernizzi, G., Lambrughi, M., Regonesi, M.E., Tortora, P., Papaleo, E. (2013).
Biochim. Biophys. Acta, 1830 (11):5236-47