UNIVERSITA' DEGLI STUDI DI MILANO-BICOCCA

Facoltà di Scienze Matematiche, Fisiche e Naturali

Scuola di Dottorato di Scienze

Dottorato di Ricerca in Biotecnologie Industriali

XXV ciclo

Conformational transitions of the intrinsically disordered protein Sic1 from the yeast *Saccharomyces cerevisiae*. Towards structural and functional characterization of the inhibitory complex with Cdk1-Clb5

Lorenzo Testa

Tutor: Dott. ssa Stefania Brocca

Co-tutor: Prof. Ssa Rita Grandori

Dottorato di Ricerca in Biotecnologie Industriali
XXV ciclo

Dottorando: Lorenzo Testa

Matricola: 074521

Tutor: Dott. ssa Stefania Brocca

Co-tutor: Prof. ssa Rita Grandori

Coordinatore: Prof. Marco Vanoni

Università degli studi di Milano-Bicocca

Piazza dell'Ateneo Nuovo, 1, 20126, Milano

Dipartimento di Biotecnologie e Bioscienze

Piazza della Scienza, 2, 20126, Milano

TABLE OF CONTENTS

ABSTRACT

Many naturally occurring proteins have been shown to lack rigid three-dimensional structure, existing instead as dynamic ensembles of inter-converting conformations. Many of these proteins acquire an ordered structure only upon binding to specific intracellular partners. In isolation, these proteins exhibit a highly dynamic structure that is resembling more the denatured rather than native state of "normal" globular proteins. Intrinsically disordered proteins (IDPs) have attracted a great deal of interest since it became clear that their lack of structural specificity is of physiological importance. Indeed, structural disorder characterizes a broad class of regulatory proteins, which all share the feature to bind multiple interactors. Intrinsic disorder is a common property of many protein types especially among the proteins involved in signal transduction and transcription regulation. The dynamic properties of IDPs are considered to be instrumental to adaptation to different interaction surfaces and to favor rapid formation and dissociation of the complexes, as required for efficient intracellular regulatory networks.

In the yeast *Saccharomyces cerevisiae*, Sic1 is a central regulatory protein of the cell cycle, acting as inhibitor of the cyclin-dependent protein kinases by forming ternary complexes with kinases and their cognate cyclins. A well established role of Sic1 is to control the timing of entrance in S phase during the yeast cell cycle by inhibiting the complex Cdk1-Clb5/6, whose activity is required for the G1-to-S transition. The inhibition is released upon ubiquitin-dependent Sic1 degradation, which is, in turn, triggered by Sic1 phosphorylation at multiple sites in its N-terminal region. The Sic1 functional kinase-inhibitor domain (KID) corresponds to the C-terminal 70, out of 284, amino acids forming the native protein. This Sic1 C-terminal region is structurally and functionally related to the mammalian p21 and p27 KIDs that are located, instead, at the protein N-terminus. A detailed structural characterization of the mammalian proteins have been achieved, both on their unbound states and on the complexes with their partners.

Previous work has shown that Sic1 is disordered in its whole length, with some intrinsic propensity to form ordered helical structure.

Structural characterization of proteins in disordered conformation is technically difficult, but it is important to better understand folding transitions to ordered states. In this work, an in-depth description of Sic1 in the absence of interactors, both in terms of secondary and tertiary structure, is presented. A multiparametric analysis, which employed a set of complementary methods sensitive to distinct structural features, has been

used to guarantee the description of such a highly dynamic and heterogeneous molecular ensemble.

First of all, a novel tool for extracting structural information from electrospray-ionization mass spectrometry (ESI-MS) data has been developed. It has been shown that the extent of protein ionization under nondenaturing conditions correlates with the solvent-accessible surface area (SASA), for either folded or unfolded proteins. Therefore, the technique has been employed to estimate the SASA of Sic1.

Fragments corresponding to the N- and C-terminal moieties of Sic1 have been produced, and circular dichroism (CD) data showed that the little content in secondary structure is distributed quite uniformly throughout the chain length, although the C-terminus is slightly more ordered than the N-terminus. Consistent with such evidence, conformational analysis by ESI-MS suggested that the Sic1 C-terminal domain is more structured than its complementary N-terminal domain. Thus, altogether, functional and structural features pointed to a modular organization of this protein, despite its disordered nature.

A more detailed description of the Sic1 KID has been achieved by integrating an array of biophysical data with all-atom molecular dynamics (MD) simulations. Highly dynamic helical elements are detected by Fourier-transformed infrared (FT-IR) spectroscopy, while protein tertiary structure has been probed by nondenaturing gel filtration, ESI-MS, and electrospray-ionization ion-mobility mass-spectrometry (ESI-IM-MS). The molecular ensemble of the isolated KID fragment has been found to interconvert between collapsed states of different compactness, with a small fraction of the population found in a highly compact state. MD simulations results has suggested a predominant role of electrostatic interactions in promoting the compaction in the Sic1 inhibitory domain. Moreover, comparison to the full-length protein hinted to a critical role of chain length in determining the overall compaction of Sic1.

Since no structural data are available for the Sic1-Cdk1-Clb5/6 ternary complex, the molecular mechanism by which Sic1 inhibits S-Cdk1 activity remains unclear. In this work, some preliminary results about the heterologous expression and purification of Cdk1 and Clb5 are also presented. Cdk1 has been overexpressed as a histidine-tagged protein in *Escherichia coli* cells and CD analysis has revealed a similar fold to mammalian homologues. The full-length Clb5, on the contrary, underwent proteolytic degradation when expressed in *E. coli* cells, probably because of the high degree of disorder in the N-terminal region of the amino-acids sequence. The deletion of the first 156 residues has allowed the expression

of the protein as inclusion bodies (IBs). A purification procedure has been then developed, so that IBs have been solubilized and refolded by on-column removal of denaturing conditions. The preliminary results concerning the expression and the characterization of Cdk1 and Clb5 represent a starting points for the production of a functional Cdk1-Clb5 complex and so, ultimately, for the complete description of Sic1 binding mechanism.

RIASSUNTO

E' stato ampiamente dimostrato che molte proteine sono prive di una rigida struttura tridimensionale, bensì esistono come un *ensemble* di conformazioni che interconvertono tra loro in modo dinamico. Molte di queste proteine acquisiscono una struttura ordinata solo in concomitanza con il legame a specifici interattori intracellulari. Nella loro forma isolata, queste proteine esibiscono una struttura altamente dinamica che può essere accomunata più allo stato denaturato che al "normale" stato nativo delle proteine globulari. Le proteine intrinsecamente disordinate (*intrinsically disordered proteins*, IDPs) hanno attratto sempre maggiore interesse da quando è stato dimostrato che la loro mancanza di specificità strutturale è importante dal punto di vista fisiologico. Infatti, il disordine strutturale caratterizza un'ampia classe di proteine regolatorie, il cui ruolo prevede il legame con interattori multipli. Il disordine intrinseco è una proprietà cha accomuna molti tipi di proteine, specialmente quelle coinvolte nella trasduzione del segnale e nella regolazione della trascrizione. La natura dinamica della IDPs è considerata funzionale all'adattamento verso differenti superfici di interazione e favorisce la rapida formazione e dissociazione di complessi multiproteici. Queste caratteristiche sono necessarie per instaurare efficienti networks regolatori all'interno della cellula.

Nel lievito *Saccharomyces cerevisiae*, Sic1 è una proteina con un ruolo centrale nella regolazione del ciclo cellulare. La proteina agisce come inibitore delle chinasi ciclina-dipendenti, formando un complesso ternario con le chinasi e le cicline. In particolare, Sic1 controlla l'ingresso nella fase S durante il ciclo cellulare del lievito inibendo i complessi Cdk1-Clb5/6, la cui attività è necessaria per la transizione G1/S. L'inibizione viene annullata in seguito alla degradazione ubiquitina-dipendente di Sic1. Questa degradazione, a sua volta, è innescata da eventi multipli di fosforilazione nella regione N-terminale della proteina.Dei 284 residui della catena amminoacidica di Sic1, i 70 amminoacidi della regione C-terminale corrispondono al dominio avente funzione inibitoria (*kinase-inhibitor domain*, KID). Questa regione è strutturalmente e funzionalmente correlata ai KIDs delle proteine di mammifero p21 e p27. In questo caso, tuttavia, i domini inibitori sono posti nella regione N-terminale della sequenza. Una caratterizzazione dettagliata delle proteine di mammifero, sia in assenza dei propri interattori sia in complesso con essi, è stata effettuata.

Un precedente lavoro ha indicato che Sic1 è disordinata lungo tutta la sua lunghezza, ma possiede un'intrinseca propensione a formare strutture ad elica.

La caratterizzazione strutturale di proteine nella conformazione disordinata è difficoltosa dal punto di vista tecnico, ma è importante per individuare nel dettaglio le transizioni che portano alla formazione di strutture ordinate. In questo lavoro viene presentata un'approfondita descrizione di Sic1 in assenza di interattori, in termini di struttura secondaria e terziaria. Per garantire la completa descrizione di un sistema molecolare così dinamico ed eterogeneo è stata effettuata un'analisi multiparametrica, in cui è stata impiegata una serie di metodi in grado di definire differenti caratteristiche strutturali.

Innanzitutto, si è sviluppato un innovativo metodo per estrarre informazioni strutturali da dati di spettrometria di massa a ionizzazione elettrospray (*electrospray-ionization mass spectrometry*, ESI-MS). E' stato dimostrato che, sia per quanto riguarda proteine globulari che proteine denaturate, il grado di ionizzazione correla con la superficie accessibile al solvente (*solvent-accessible surface area*, SASA) delle proteine stesse. Perciò, la tecnica è stata utilizzata per stimare la SASA di Sic1.

Frammenti corrispondenti alle regioni ammino- e carbossiterminale di Sic1 sono stati prodotti, e i dati di dicroismo circolare (*circular dichroism*, CD) hanno indicato che il piccolo contenuto di struttura secondaria è distribuito piuttosto uniformemente lungo tutta la sequenza, sebbene la regione carbossiterminale sia leggermente più ordinata di quella amminoterminale. Successive analisi conformazionali effettuate mediante ESI-MS hanno confermato quest'evidenza, suggerendo che il dominio carbossiterminale sia più strutturato del complementare dominio amminoterminale. Quindi sia le caratteristiche funzionali che strutturali hanno messo in evidenza l'organizzazione modulare di questa proteina, nonostante la sua natura disordinata.

Una descrizione più dettagliata del KID di Sic1 è stata effettuata mediante l'integrazione di un insieme di evidenze di carattere biofisico con simulazioni di dinamica molecolare (*molecular dynamics*, MD). Mediante analisi di spettroscopia infrarossa a trasformata di Fourier (*Fourier-transformed infrared*, FT-IR) sono stati individuati elementi ad elica altamente dinamici, mentre la struttura terziaria della proteina è stata saggiata mediante gel filtrazione analitica, ESI-MS e spettrometria di massa a mobilità ionica (*electrospray-ionization ion-mobility mass-spectrometry*, ESI-IM-MS). Si è verificato che l'ensemble molecolare del KID nella sua forma isolata è formato da stati collassati di differente compattezza che interconvertono tra di loro, con una piccola frazione della popolazione che si trova in uno stato altamente compatto. Le simulazione computazionali hanno poi suggerito un ruolo predominante delle interazioni elettrostatiche nella

compattazione del dominio inibitore di Sic1. Inoltre, il confronto con la proteina integra ha messo in evidenza anche un ruolo critico della lunghezza della catena amminoacidica nella determinazione del grado totale di compattezza della proteina.
Dal momento che non sono disponibili dati di carattere strutturale per quanto riguarda il complesso ternario Sic1-Cdk1-Clb5/6, il meccanismo molecolare mediante il quale Sic1 inibisce l'attività del complesso chinasico rimane non chiaro. In questo lavoro vengono presentati anche I risultati relativi all'espressione eterologa e alla purificazione di Cdk1 e Clb5. Cdk1 è stata overespressa come proteina recante un tag di istidine, in cellule di *Escherichia coli,* e le successive analisi CD hanno rivelato che la proteina assume un ripiegamento simile a quello degli omologhi di mammifero. Al conrario, Clb5, nella sua forma integra, è andata incontro a degradazione proteolitica quando espressa in cellule di E. coli, probabilmente a causa dell'elevato grado di disordine posseduto dalla regione amminoterminale della sequenza. La delezione dei primi 156 amminoacidi ha permesso l'espressione della proteina sottoforma di corpi di inclusione (*inclusion bodies*, IBs). E' stata quindi sviluppato un procedimento di purificazione mediante il quale gli IBs sono stati solubilizzati e la proteina è stata posta in condizioni native allontanando il denaturante dalla colonna cromatografica.
I risultati preliminari relativi all'espressione e alla caratterizzazione di Cdk1 e Clb5 rappresentano un importante punto di partenza per la produzione di un complesso Cdk1-Clb5 funzionale e, in ultima analisi, per la completa descrizione del meccanismo inibitorio di Sic1.

1. INTRODUCTION

1.1 Intrinsically Disordered Proteins

Proteins are the major components of the living cell. They play crucial roles in the maintenance of life and protein dysfunctions may cause development of various pathological conditions. For a very long time it has been believed that the specific functionality of a given protein is predetermined by its unique tridimensional structure (Lemieux and Spohr, 1994). It is recognized now that the fate of any given polypeptide chain is determined by the peculiarities of its amino acid sequence. Although many proteins are predisposed to fold into unique structures which evolved to possess unique biological functions, they can misfold either spontaneously or due to the mutations and other genetic alterations, problematic processing or posttranslational modifications, or due to the exposure to harmful environmental conditions (Figure 1). For more than five decades, individual proteins that possess no definite ordered tridimensional structure but still play important biological roles have been identified. The discovery rate for such proteins has been increasing continually and has become especially rapid during the last decade (Dunker *et al.*, 2008a). Such proteins are widely known as intrinsically disordered proteins (IDPs) because they exist as structural ensembles, either at the secondary or at the tertiary level, under physiological conditions. Both extended regions with some elements of secondary structure and collapsed (molten globule-like) domains with poorly packed side chains are included in the definition of intrinsic disorder. The discovery and characterization of these proteins is becoming one of the fastest growing areas of protein science and it is recognized now that many such proteins with no unique structure have important biological functions (Radivojac *et al.*, 2007; Dunker *et al.*, 2008b)

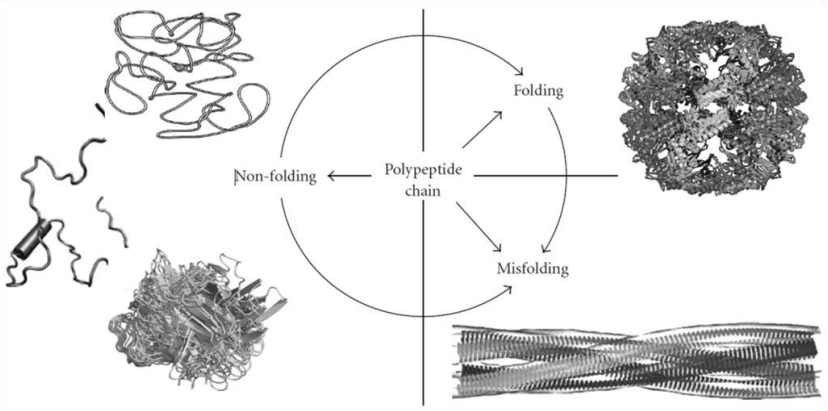


Figure 1. Fates of a polypeptide chain. Left, three structures representing typical IDPs with different levels of disorder (from top to bottom): native coil, native premolten globule, and native molten globule. Right, top structure illustrates a natively-folded protein, whereas the bottom structure represents one of the products of protein misfolding—a molecular model of the compact, 4-protofilament insulin fibril. From (Uversky, 2010)

In general, an IDP function either directly stems from the protein's ability to fluctuate over an ensemble of structural states, or it is realized via binding to one or several partner molecule(s) in a structurally adaptive process (Figure 2). These functional capacities are exploited in many molecular settings and thus IDPs may fulfill many different functions (Dunker *et al.*, 2002). Predictions in various functional classes of proteins have shown that disorder is primarily associated with signal transduction, cell-cycle regulation and gene expression (Ward *et al.*, 2004). Intriguingly, IDPs were shown to be involved in different diseases, even cancer, giving rise to the "disorder in disorders" or D^2 concept (Uversky *et al.*, 2008; Iakoucheva *et al.*, 2002). The crucial role of IDPs in signaling is further confirmed by the fact that eukaryotic proteomes, with their extensively developed interaction networks, are highly enriched in unstructured proteins, relative to bacteria and archaea. Indeed, putative, long (>30 residue) disordered segments were found to occur in 2.0% of archaean, 4.2% of eubacterial and 33.0% of eukaryotic proteoms (Ward *et al.*, 2004). Moreover, application of a data mining tool to over 200000 proteins from Swiss-Prot database revealed that many protein functions are associated with long disordered regions (Xie *et al.*, 2007). Indeed, of the 711 Swiss-Prot functional keywords associated with at least 20 proteins, 262 were found to be strongly positively correlated with long intrinsically disordered regions (IDRs), whereas 302 were strongly negatively correlated with such regions (Xie *et al.*, 2007). Therefore, the functional diversity provided by disordered regions complements functions of ordered protein domains.

By considering the mechanistic details of their various modes of action, the many different functions of IDPs can be divided into six general categories (Figure 2) (Tompa, 2002). The first general functional class of IDPs is that of *entropic chains*, whose function stems directly from their ensemble of structural states of similar conformational energies. These proteins, subclassified as entropic springs, bristles/spacers or linkers, either generate force against structural changes or influence the orientation/localization of attached domains (Dunker *et al.*, 2002). In the other five classes, IDPs function via molecular recognition, i.e. they permanently or transiently bind another macromolecule or small ligand(s). The function known as *display*

sites has a role in post-translational modifications. Indeed, the action of a modifying enzyme requires flexibility of the substrate, which enables transient but specific interaction with the active site of the enzyme. Pertinent to this function are, for instance, the success of disorder-based prediction of phosphorylation sites (Iakoucheva *et al.*, 2004) and the evidence of disorder-targeted ubiquitination of securin and cyclin B in cell cycle regulation (Cox *et al.*, 2002; Prakash *et al.*, 2004). IDPs, by transiently binding to their partners, can also act as *chaperones* (Tompa and Csermely, 2004). It has been found that RNA chaperones have a much higher incidence of disorder than any other functional class: 40% of their residues falls into long disordered regions (>30 residues), whereas the same number is 15% for protein chaperones (Tompa and Csermely, 2004). Further, the function of many, or possibly all, of these proteins depends directly on disorder in a way that the disordered segment serves for either recognizing, solubilizing or loosening the structure of the misfolded ligand. To account for these mechanistic details, an entropy transfer model of disorder in chaperone function has been suggested. According to this model, the ordering of the chaperone is accompanied by the concomitant unfolding of the substrate (Tompa and Csermely, 2004). Disordered proteins that function by permanent binding to their partners belong to the classes of *effectors*, *assemblers* and *scavengers*. *Effectors* bind and modify the activity of their partners (Tompa, 2002). Their action is mostly inhibitory, but they may also activate another protein, demonstrating their extreme structural and functional versatility. The effector protein $p21^{Cip1}$, for example, has been shown not only to inhibit cyclin-dependent kinases (Cdks), but also to be able to assemble the cyclin-Cdk complex leading to Cdk activation (Olashaw *et al.*, 2004) (see Section 1.3). IDPs classified as *assemblers* regulate the formation of multiprotein complexes and/or target the activity of attached domains. Such proteins have been identified in the assembly of ribosome, cytoskeleton, transcription pre-initiation complex and chromatin (Tompa, 2002). Finally, *scavengers* store and/or neutralize small ligands. The classical examples of this mode of action are casein(s), which prevent calcium phosphate precipitation in the milk by capturing small seeds as soon as they form, or salivary proline-rich glycoproteins, which form tight complexes with tannins that can resist harsh conditions encountered in the digestive tract (Tompa, 2002). It is important to note that the previously described functional modes are not exclusive. Different domains within the same protein, or even the same amino-acids sequences, may be involved in distinct functional roles. For example, securin, as an effector, is an inhibitor

of separase, but it is also required for the activation of the enzyme via a chaperone-like action (Uhlmann, 2003).

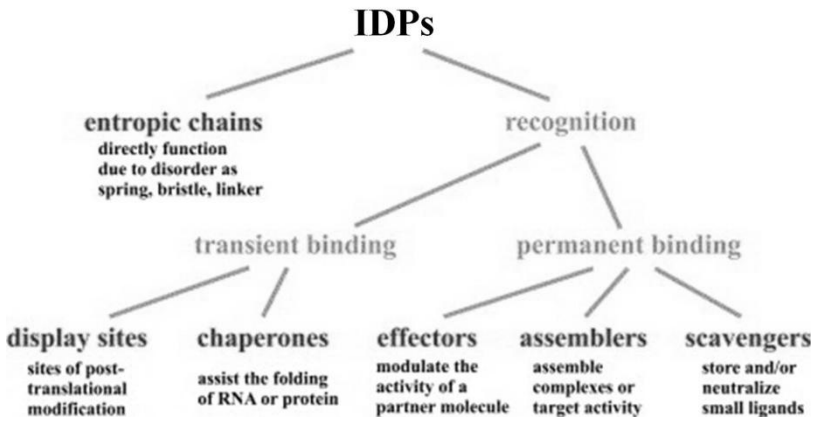


Figure 2. Functional classification scheme of IUPs. The function of IUPs stems either directly from their capacity to fluctuate freely about a large configurational space (entropic chain functions) or ability to transiently or permanently bind partner molecule(s). For each functional class, a short definition of function is given. From (Tompa, 2005).

The number of IDPs and IDRs in various proteomes is very large; for instance, in mammals, ~75% of the signaling proteins are predicted to contain long disordered regions (>30 residues), about half of the total proteins are predicted to contain such long disordered regions, and ~25% of the total proteins are predicted to be fully disordered. IDPs and IDRs have amazing structural variability and possess a very wide variety of functions. For all these reasons the *unfoldome* and *unfoldomics* concepts were recently introduced (Cortese *et al.*, 2005). The term *unfoldome* refers to the portion of proteome which includes IDPs. *Unfoldomics* considers not only the identity of disordered proteins of a given organism but also their functions, structures, interactions and evolution (Uversky *et al.*, 2009).

It is clearly recognized now that the disorderedness is linked to the peculiarities of amino acid sequences, as IDPs/IDRs exhibit low sequence complexity and are generally enriched in polar and charged residues and are depleted of hydrophobic residues (other than proline). In particular, IDPs are enriched in disorder-promoting amino acids (A, R, G, Q, S, P, E and K) and depleted in order-promoting amino acids (W, C, F, I, Y, V, L and N) (Dunker *et al.*, 2001). These features result in their inability to fold into globular structures and form the basis of computational tools for disorder prediction (Dosztanyi and Tompa, 2008). These same computational tools

can also be utilized for the large-scale discovery of IDPs in various proteomes.

IDPs possess a number of very distinctive structural properties. These feature include, but are not limited to, sensitivity to proteolysis (Hubbard *et al.*, 1998), aberrant migration during SDS-PAGE (Iakoucheva *et al.*, 2001) and insensitivity to denaturing conditions (Reeves and Nissen, 1999). Definitive disorder characteristics can then be defined by hydrodynamic measurement, circular dichroism (CD) spectropolarimetry, infrared (IR) and nuclear magnetic resonance (NMR) spectroscopy, mass spectrometry (MS) and Molecular Dynamics (MD) simulation (see Section 1.4).

IDPs range from completely unstructured polypeptides to extended partially structured forms to compact disordered ensembles containing substantial secondary structure (Uversky, 2002a; Dunker *et al.*, 2008a). They exist as dynamic ensembles, also known as "protein clouds" (Dunker and Uversky, 2010), in which the atom positions and backbone Ramachandran angles vary significantly over time with no specific equilibrium values. Typically, these positions and angles undergo non-cooperative conformational changes. Although these protein clouds are highly dynamic, often their structures can be described rather well by a fairly limited number of lower-energy conformations (Huang and Stultz, 2008). As a result, both extended regions (random coil-like) as well as partially collapsed regions with residual secondary structure (pre-molten globule-like), domains with noticeable secondary structure and compaction (partially folded or molten globule-like), and domains with poorly packed side chains are included in the current view of intrinsic disorder (Figure 1) (Uversky, 2002a). It has been shown that the range of conformational changes induced in natively unfolded proteins as a result of their interaction with natural partners is very wide (Uversky, 2002a). In fact, examples of all possible conformational transitions have been described, including function- induced transitions of coil to pre-molten globule, coil to molten globule, coil to ordered conformation, pre-molten globule to molten globule, pre-molten globule to rigid structure and molten globule to ordered, and rigid form (Uversky, 2002b). This means that the structure-function paradigm, which emphasizes that ordered 3D structures represent an indispensable prerequisite to effective protein functioning, should be redefined to include intrinsically unstructured proteins (Dunker *et al.*, 2001). According to this redefined paradigm, native proteins (or their functional regions) can exist in any of the known conformational states, ordered, molten globule, pre-molten globule and coil. Function can arise from any of these conformations and transitions between them. Thus, not

just the ordered state but any of the known polypeptide conformations can be the native state of a protein.

Modularity of intrinsic disorder

Many proteins contain mixtures of ordered and disordered regions, with the latter frequently appearing as independent functional units. It has been shown, indeed, that long disordered regions in proteins often have the same characteristics of modules in multidomain proteins (Pentony and Jones, 2010). There is an increasingly strong evidence that long disordered regions can represent another type of protein domain and not only flexible spacer regions. Many evidence of across-species conservation for a significant fraction of disordered regions have been shown (Pentony and Jones, 2010). Moreover, it has been found that, after alternative-splicing events, disordered regions are either entirely maintained or entirely removed (Pentony and Jones, 2010; Romero *et al.*, 2006). These features are the same observed for other "classic" protein domains, supporting the hypothesis of a reclassification of certain disordered regions as a novel type of protein domain. There are three operative definitions for a "classical" protein domain: i) an autonomous structural unit of a protein; ii) a protein segment that can be recognized in distinct genetic contexts by virtue of sequence similarity; iii) an interchangeable element of a protein with functional autonomy (Tompa *et al.*, 2009). It is important to note that, often, these definitions do not hold simultaneously. Some disordered regions serve as functional elements, and, although it may seem perplexing, they are also structurally autonomous, because they can be taken out of context while preserving their structural state and function. Given all of the above, it has been proposed that disordered domains are a distinct class of protein architecture (Tompa *et al.*, 2009). To clarify whether a given protein or region belongs to this novel class, disordered domains have been defined as lengthy regions that display the following four characteristics: (1) they are structurally and functionally independent of the remainder of the protein molecule (or they constitute the entire protein); (2) they can be recognized by homology due to evolutionary conservation of sequence; (3) they are structurally characterized and/or predicted as disordered; and (4) they possess at least one specific biological function (Tompa *et al.*, 2009). Actually, regions with several of these characteristics (and sometimes with all four) have been termed domains in the literature for a long time, without reference to their actual structural status. With the advent of the concept of

protein disorder, the disordered domain ought to be recognized as a distinct element of protein architecture.

Binding mechanisms of IDPs and functional misfolding

Unbound IDPs are disordered in solution and often perform their biological functions by binding to their specific partners. Folding of disordered proteins into ordered structures may occur upon binding to their specific partners (Sugase *et al.*, 2007; Gunasekaran *et al.*, 2003) which may allow disordered regions to structurally accommodate multiple interaction partners with high specificity and low affinity (Demchenko, 2001; Tompa *et al.*, 2005). Anyway, it has become increasingly clear that some IDPs retain significant disorder even in their complexes (Mittag *et al.*, 2008; Sigalov *et al.*, 2007; Tompa and Fuxreiter, 2008).

The recognition function of IDPs can be realized via several molecular mechanisms, being frequently associated with the disorder-to-order transition induced by binding to their partners. The binding-coupled folding of IDPs/IDRs may be induced by the template or be selected from the ensemble of conformations (Uversky, 2011). In other words, the IDP structure adopted in the bound form may be enforced by the partner molecule or reflect the inherent conformational preferences of IDPs. One of the models for finding intrinsic disorder-based interactors, Molecular Recognition Feature (MoRF) model, involves a short binding region located within a longer disordered region (Oldfield *et al.*, 2005). MoRFs were proposed from the study of existing protein complex structures, and application of the MoRF model to proteomes suggests that MoRFs may be a common mediator of protein-protein interactions (Oldfield *et al.*, 2005). Alternative models of MoRF-like interactions are the short linear motif (SLiM) or eukaryotic linear motif (ELM) based on sequence motifs that are recognized by peptide recognition domains (Puntervoll *et al.*, 2003). A different approach is taken by the ANCHOR model, which identifies segments of disordered regions that are likely to fold in conjunction with a globular binding partner (Dosztanyi *et al.*, 2009). In the primary contact site (PCS) model, certain regions within the disordered ensemble are more exposed than others, and thereby may serve as the first sites of contact with the partner (Csizmok *et al.*, 2005). Some IDPs in the unbound state were proposed to have strong conformational preferences for their bound conformations; i.e., they use partially/transiently pre-formed elements for recognition, also defined as intrinsically folded structural units (IFSUs) (Fuxreiter *et al.*, 2004; Sivakolundu *et al.*, 2005). Although the concepts

underlying the above described interaction models are closely related, some important differences have to be mentioned. Indeed, a PCS/anchor site can be defined, in kinetic terms, as a recognition element that forms the initial contact with the partner, whereas a MoRE represents, from a thermodynamic point of view, a region in the molecular interface that contributes the major part of the free energy of binding. A more detailed description of the recognition mechanism has been achieved by the "fly-casting" model, which suggests that IDPs make use of their folding funnel in binding to their partner (Shoemaker *et al.*, 2000). According to this mechanism, a dimensionality reduction occurs when the folding of a disordered protein is coupled with binding, thereby speeding up the search for specific targets. This model invokes both the greater capture radius of IDPs and the mechanistic coupling of the recognition process to folding, in which pre-formed, exposed, recognition elements may be effective mechanistic devices.

As already stated, very often, IDPs lack the hydrophobic cores typical for ordered proteins and cannot be described as single, rigid structures clearly resembling instead highly dynamic hairballs or diffuse protein clouds. However, even these apparently unordered clouds might have some local preferences for transient secondary structure elements and even for some transient tertiary contacts. Such dynamic pre-organization imposes spatial restrictions on IDPs, therefore exposing some of their potential contact sites. The existence of such pre-formed binding sites enables faster and more effective interactions of IDPs with their targets (Uversky and Dunker, 2010). Despite their highly flexible structure combined with the interspersed preformed dynamic binding elements possessing high intrinsic potential to interact, IDPs are typically able to escape unwanted interactions with non-native partners. Surprisingly, more and more data support the idea that IDPs do so via the so-called functional misfolding (Uversky, 2011). According to this model, the mentioned sticky preformed secondary structure elements are dynamically involved in a set of intramolecular non-native interactions. As a result of such functional misfolding, the interaction-prone preformed fragments are sequestered inside a less-interactive cage, where they are dynamically excluded from the environment and therefore might escape unwanted interactions with the non-native binding partners. Both non-native intramolecular electrostatic and hydrophobic interactions might contribute to the functional misfolding. Although the currently available information about structures of unbound IDPs in solution and about their long-range intramolecular interactions is very sparse, the existence of functional misfolding has been illustrated for several IDPs (Marsh *et al.*, 2010; Kussie

et al., 1996; Slep *et al.*, 2001). Based on simple biophysical considerations, there is a good chance that this phenomenon is highly abundant at least among the native-coil like and native pre-molten globule-like IDPs. Therefore, this functional misfolding, which sequesters and thus preserves interaction-prone elements, is in a sharp contrast to the pathological protein misfolding; i.e., a process by which protein molecules adopt an aggregation-prone conformation that can lead to amyloids formation (Figure 1). Protein misfolding is believed to play a crucial role in the development of a huge variety of unrelated diseases, such as prion diseases, Alzheimer's disease, Parkinson's disease, diabetes and cancer, and many others, that share the pathological feature of aggregated misfolded protein deposits.

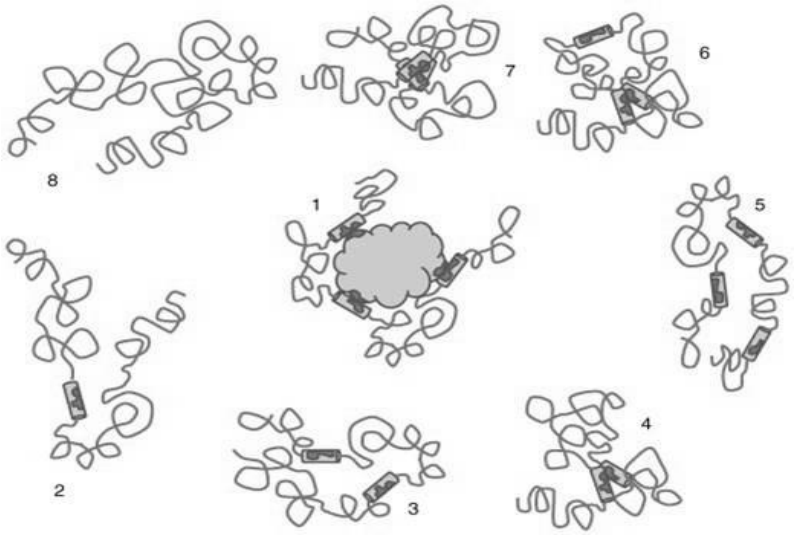


Figure 3. Model illustrating the concept of functional misfolding. Structure of an IDP bound to its native partner is shown at the middle of the figure (state 1). Other structures are illustrative members of a dynamic ensemble of conformations probed by the unbound IDP. Various, partially condensed states of preformed secondary-structure elements with non-polar surfaces and different less condensed states are shown. Different partially folded states will be populated with different probability. For example, states 3 and 5 are probably unfavorable in this ensemble, since all their hydrophobic surfaces are solvent-exposed. State 2, instead, has its hydrophobic patch partially protected from the solvent by the loop. States 4, 6, and 7 have at least part of their hydrophobic patches involved in intrachain interactions. Comparison of the structures attainable by the unbound IDP with its bound structure clearly illustrates the functional misfolding principle: formation of interaction-prone structural elements inevitably leads to the formation of non-native long-range intramolecular interactions. From (Uversky, 2011).

1.2 The eukaryotic cell cycle

Cellular division is a complex process in which all cellular components are duplicated and carefully distributed to daughter cells. An elaborate regulation system evolved to guarantee that each synthesis and localization event occurs at the proper time and in the proper space, also responding to messages coming from the extracellular environment. The coordination among different events is mediated, to a large extent, by the control on the activity and the localization of cyclin-dependent protein kinases (Cdks).

Cdks are a family of serine/threonine protein kinases whose members are small proteins (~34–40 kDa) composed of little more than the catalytic core shared by all protein kinases. By definition, all Cdks share the feature that their enzymatic activation requires the binding of a regulatory cyclin subunit. Cyclins belong to a large family of proteins which accumulate during interphase of the eukaryotic cell cycle and are destroyed at the conclusion of mitosis. They generally function as enzyme activators that drive the cell-cycle progression. In most cases, full activation of Cdks also requires phosphorylation of a threonine residue near the kinase active site (Morgan, 2007).

Although originally identified as enzymes that control cell-cycle events, members of the Cdk family are involved in other cellular processes as well. Mammalian cells, for example, contain at least nine Cdks, only four of which (Cdk1, 2, 4 and 6) are involved directly in cell-cycle control (Table 1). Another family member (Cdk7) contributes indirectly by acting as a *Cdk-activating kinase* (*CAK*) that phosphorylates other Cdks. Cdks are also components of the machinery that controls basal gene transcription by RNA polymerase II (Cdk7, 8 and 9) and are involved in controlling the differentiation of nerve cells (Cdk5).

In the fission yeast *Schizosaccharomyces pombe* and the budding yeast *Saccharomyces cerevisiae*, all cell-cycle events are controlled by a single essential Cdk called Cdk1. Cell-cycle events in multicellular eukaryotes are controlled by two Cdks, known as Cdk1 and Cdk2, which operate primarily in M phase and S phase, respectively. Mammalian cells also contain two Cdks (Cdk4 and Cdk6) that are important in regulating entry into the cell cycle in response to extracellular factors.

Cdk function has been remarkably well conserved during evolution. It is possible, for example, for yeast cells to proliferate normally when their gene for Cdk1 is replaced by the human one. This and other evidence clearly illustrate that Cdk function, and thus the function of the cell-cycle control

system, has remained fundamentally unchanged over hundreds of millions of years of eukaryotic evolution (Liu and Kipreos, 2000).

Cdks exert their effects on cell-cycle events by phosphorylating a large number of proteins inside the cell. During mitosis in particular, when many aspects of cellular architecture and metabolism are altered, Cdks phosphorylate hundreds of distinct proteins. These Cdk substrates are phosphorylated at serine or threonine residues in a specific sequence context that is recognized by the active site of the Cdk protein. In most cases, the target serine (S) or threonine (T) residue is followed by a proline (P); it is also highly favorable for the target residue to have a basic amino acid two positions after the target residue. The typical phosphorylation sequence for Cdks is [S/T*]PX[K/R], where S/T* indicates the phosphorylated serine or threonine, X represents any amino acid and K/R represents the basic amino acid lysine (K) or arginine (R).

Species	Name	Original Name	Size (amino acids)	Function
S. cerevisiae	Cdk1	Cdc28	298	All cell-cycle stages
S. pombe	Cdk1	Cdc2	297	All cell-cycle stages
D. melanogaster	Cdk1	Cdc2	297	M
	Cdk2	Cdc2c	314	G1/S, S, possibly M
	Cdk4	Cdk4/6	317	G1, promotes growth
X. laevis	Cdk1	Cdc2	301	M
	Cdk2		297	S, possibly M
H. sapiens	Cdk1	Cdc2	297	M
	Cdk2		298	G1/S, S, possibly M
	Cdk4		303	G1
	Cdk6		326	G1

Table 1. Cdks that control the cell cycle. From (Morgan, 2007).

All protein kinases have a tertiary structure comprising a small amino-terminal lobe and a larger carboxy-terminal lobe (De Bondt *et al.*, 1993). ATP fits snugly in the cleft between the lobes, in such a way that the phosphates are oriented outwards, toward the mouth of the cleft. The protein substrate binds at the entrance of the cleft, interacting mainly with the surface of the carboxy-terminal lobe. Nearby residues catalyze the transfer of the terminal □-phosphate of ATP to a hydroxyl oxygen in the protein substrate.

Cdks have the same two-lobed structure as other protein kinases, but with two modifications that make them inactive in the absence of cyclin. These modifications have been revealed by detailed crystallographic studies of the structure of human Cdk2 (De Bondt *et al.,* 1993). First, a large, flexible loop, the T-loop or activation loop, rises from the carboxy-terminal lobe to block the binding of protein substrate at the entrance of the active-site cleft. Second, in the inactive Cdk, several important amino-acid side chains in the active site are incorrectly positioned, so that Cdk activation requires extensive structural changes in the active site. Indeed, although Cdks strongly bind ATP, they remain in a catalytically inactive conformation until the binding to the cyclin occurs. This interaction induces structural rearrangements that result in the selective reorientation of ATP, correctly positioning the phosphate groups for the kinase reaction (Heitz *et al.*, 1997). Two alpha helices make a particularly important contribution to the control of Cdk activity. The highly conserved PSTAIRE helix of the amino-terminal lobe (also known as the α1 helix) interacts directly with cyclin and moves inward, toward the catalytic cleft, upon cyclin binding, causing the reorientation of residues that interact with the phosphates of ATP. The small (five-residue long) L12 helix, just before the T-loop in the primary sequence, changes structure to become a beta strand upon cyclin binding, also contributing to the reconfiguration of the active site and the T-loop.

Cyclins are the key determinants of Cdk activity and can be classified in four groups

Cyclins are a diverse family of proteins whose defining feature is that they bind and activate members of the Cdk family. Most cyclins display dramatic changes in concentration during the cell cycle, which help to generate the oscillations in Cdk activity that form the foundation of the cell-cycle control system (Evans *et al.,* 1983). The regulation of cyclin concentration, primarily by changes in cyclin gene expression and cyclins proteolysis, is therefore of fundamental importance in cell-cycle control. Cyclins, like their Cdk partners, are involved in a number of processes other than cell-cycle control (Ubersax *et al.*, 2003). Those cyclins that directly regulate Cdks controlling cell-cycle progression can be divided into four classes, based primarily on the timing of their expression and their functions in the cell cycle. Three of these classes, the G1/S cyclins, S cyclins and M cyclins, are directly involved in the control of cell-cycle events (Morgan, 1997). The fourth class, the G1 cyclins, contributes to the control of cell-cycle entry in response to extracellular factors (Nurse, 1990). The G1/S cyclins (Cln1 and

Cln2 in the budding yeast *S. cerevisiae*, cyclin E in vertebrates) oscillate during the cell cycle, rising in late G1 and falling in early S phase. The primary function of G1/S cyclin-Cdk complexes is to trigger progression through the Start checkpoint and initiate the processes leading to DNA replication, principally by shutting down the mechanisms that suppress S-phase Cdk actitity in G1. G1/S cyclins also initiate other early cell-cycle events, such as duplication of the centrosome in vertebrates and its equivalent, the spindle pole body, in yeast (Ubersax *et al.*, 2003).

The rise of G1/S cyclins is accompanied by the appearance of the S cyclins (C1b5 and C1b6 in budding yeast, cyclin A in vertebrates), which form S cyclin-Cdk complexes that are directly responsible for stimulating DNA replication (Morgan, 2007). Levels of S cyclins remain high throughout S phase, G2 and early mitosis, when they help promote early mitotic events in some cell types.

M cyclins (Clb1, 2. 3 and 4 in budding yeast, cyclin B in vertebrates) appear last in the sequence, their concentration rising as the cell approaches mitosis and peaking at metaphase. M cyclin-Cdk complexes are responsible for the striking cellular changes that lead to assembly of the mitotic spindle and the alignment of sister-chromatid pairs on the spindle at metaphase. Their destruction in anaphase brings on mitotic exit and cytokinesis (Ubersax *et al.*, 2003).

The Gl cyclins, typified by Cln3 in budding yeast and cyclin D in vertebrates, help coordinate cell growth with entry into a new cell cycle and are required in many cell types to stimulate entry into a new cell cycle at the Start checkpoint. The G1 cyclins are unusual among cyclins in that their levels do not oscillate during the cell cycle, but increase gradually throughout the cycle in response to cell growth and external growth-regulatory signals. The division of cyclins into four classes is based on their behavior in the cell cycle of yeast and of vertebrate somatic cells. This is a useful simplification but is not universally applicable. The same cyclin can have different functions or timing of expression in different cell types, or may contribute to the control of more than one cell-cycle process. In the early embryos of flies and frogs, for example, where there is only S phase and M phase and no clear Start checkpoint, cyclin E levels remain high throughout the cycle and it behaves as the major S cyclin that drives DNA replication. Cyclin A, in contrast, clearly has a mitotic function in these cells and can therefore be called an M cyclin under these circumstances (Morgan, 2007).

Cvclins contain a conserved helical core

Members of the cyclin family are often quite distinct from each other in amino-acid sequence. Sequence similarity among distantly related cyclins is concentrated in a region of about 100 amino acids known as the cyclin box, which is required for Cdk binding and activation.

Despite variations in their primary structure, all cyclins are predicted to possess a similar tertiary structure known as the cyclin fold, which comprises a core of two compact domains each containing five alpha helices (Brown *et al.*, 1995). The first five-helix bundle corresponds to the conserved cyclin box. The second five-helix bundle displays the same arrangement of helices as the first, despite limited sequence similarity between the two subdomains.

Cyclin sequences are highly divergent. The length of the amino-terminal region is particularly variable, and contains regulatory and targeting regions that are specific for each cyclin class. For example, the amino-terminal regions of S and M cyclins contain short destruction-box motifs that target these proteins for proteolysis during mitosis.

The cyclin fold of twin five-helix bundles is also found in other proteins, including members of the pRB family, which regulate gene expression at the G1/S checkpoint. The cyclin fold is also found in the RNA polymerase II transcription factor, TFIIB. These structural relationships raise the intriguing possibility that cyclins and transcriptional regulators evolved from a common ancestor (Morgan, 2007).

Full Cdk activity requires phosphorylation by the Cdk-activating kinase

Cyclin binding alone is not enough to fully activate Cdks involved in cell-cycle control. Complete activation of a Cdk, and normal Cdk function in the cell, also requires phosphorylation of a threonine residue adjacent to the kinase active site (Morgan, 1997). Phosphorylation at this site is catalyzed by enzymes called Cdk-activating kinase (CAKs) (Liu and Kipreos, 2000). Phosphorylation is often thought to be a reversible modification that changes enzyme activity under particular conditions. Surprisingly, the activating phosphorylation of Cdks does not seem to behave in this way. CAK activity is maintained at a constant high level throughout the cell cycle and is not regulated by any known cell-cycle control pathway. In addition, in mammalian cells, phosphorylation can occur only after cyclin is bound, whereas in budding yeast cells phosphorylation occurs before cyclin binding (Russo *et al.*, 1996b; Harper and Elledge, 1998). In both cases, however,

cyclin binding and not phophorylation is the highly regulated, rate-limiting step in Cdk activation. Activating phosphorylation can therefore be viewed at simply a post-translational modification that is required for enzyme activity. It is not clear why the requirement for Cdk phosphorylation has been so highly conserved during evolution if it is not exploited for regulatory purposes. The constant high level of activity of CAK may, however, be explained by the fact that it has a role in transcription as well as in cell-cycle regulation (Morgan, 2007).

The identity of CAK varies dramatically in different species. In vertebrates and *Drosophila*, the major CAK is a trimeric complex containing a Cdk-related protein kinase known as Cdk7, along with its activating partner, cyclin H, and a third subunit, Mat1 (Liu and Kipreos, 2000). In budding yeast, however, CAK is a small monomeric protein kinase known as Cak1, which bears only distant homology to the Cdks. The Cdk7 homolog in budding yeast (known as Kin28) does not possess CAK activity. Fission yeast seems to be intermediate between vertebrates and budding yeast, in that it uses two CAKs: one (a complex of Mcs6 and Mcs2) that is related to the vertebrate Cdk7-cyclin H complex, and another (Csk1) that more closely resembles budding yeast Cak1.

Cdk function is regulated by inhibitory phosphorylation by Wee1 and dephosphorylation by Cdc25

Whereas the activating phosphorylation of Cdks is not regulated, two inhibitory phosphorylations do have important functions in the regulation of Cdk activity (Brown *et al.*, 1999). One is at a conserved tyrosine residue, Tyr 15 in fission-yeast Cdk1 and in human Cdks and Tyr 19 in budding-yeast Cdk1. In animal cells, additional phosphorylation of an adjacent threonine residue (Thr 14) further blocks Cdk activity. Thr 14 and Tyr 15 are located in the kinase ATP-binding site and their phosphorylation probably inhibits activity by interfering with the orientation of ATP phosphates (Jeffrey *et al.*, 1995). Changes in the phosphorylation state of these sites are particularly important in the activation of M-Cdks at the onset of mitosis, and they are also thought to influence the timing of G1/S- and S-phase Cdk activation (Nurse, 1990).

The phosphorylation state of Tyr 15 and Thr 14 is controlled by the balance of opposing kinase and phosphatase activities acting at these sites (Pavletich, 1999). One enzyme responsible for Tyr 15 phosphorylation is Wee1, which is present in all eukaryotes. In *S. cerevisiae*, the kinase is named Swe1 (Saccharomyces Wee1 homologue); fission yeast contains two kinases,

Wee1 and Mik1, which both contribute to Tyr 15 phosphorylation. Vertebrates also contain a second protein kinase, Mit1, related to Wee1, which catalyzes the phosphorylation of both Thr14 and Tyr 15. Dephosphorylation of inhibitory sites is carried out by phosphatases of the Cdc25 family, which has three members in vertebrates (one, Mih1, in the budding yeast) (Morgan, 2007; Russell *et al.*, 1989).

Wee1 and Cdc25 provide the basis for the switch-like features of M-Cdk activation, which allows abrupt and irreversible entry into mitosis. Both enzymes are regulated by their mitotic substrate, the M-phase cyclin-Cdk complex: phosphorylation by M-Cdk inhibits Wee1 and activates Cdc25. Thus, M-Cdk activates its own activator and inhibits its inhibitor, and the resulting feedback loops are thought to generate switch-like Cdk activation at the beginning of mitosis. Wee1 and Cdc25 are also important targets for regulation of Cdk activity in response to factors such as DNA damage (Pavletich, 1999).

Cdk inhibitors help suppress Cdk activity

In actively proliferating cells, most Cdk activity is suppressed during G1, resulting in stable transition period during which cell growth and other regulatory influences, such as extracellular factors, can govern entry into the next cell cycle. As discussed earlier, a combination of mechanisms suppresses Cdk activity during G1. Two of these mechanisms involve the increased cyclin destruction and decreased cyclin gene expression. The third is represented by the inhibition of Cdk activity by Cdk inhibitor proteins (CKIs) that bind and inactivate cyclin-Cdk complexes (Sherr and Roberts, 1999). These proteins are also important for promoting the arrest of the cell cycle in G1 in response to unfavorable environmental conditions or intracellular signals such as DNA damage.

Most, if not all, eukaryotic organisms possess a CKI: these include Sic1 in budding yeast, Rum1 in fission yeast and Roughex (Rux) in *Drosophila* (Table 2) (Morgan, 2007). Although these proteins display little, if any, similarity in amino-acid sequence, they share several important functional features. First, they all are potent inhibitors of the major S- and M-Cdk complexes, and all are expressed at high levels in G1 cells contributing to ensure the suppression of S- or M-Cdk activity in that phase. Second, these proteins do not inhibit G1/S-Cdks: as a result, they do not block the activation of these kinases at the Start checkpoint. Finally, these inhibitors are all targeted for destruction when phosphorylated by Cdks. In late G1,

rising G1/S-Cdk activity leads to destruction of these inhibitors and to S-Cdk activation at the beginning of S phase (Sherr and Roberts, 1999).

Given the clear importance of Sic1, Rum1 and Rux in yeast and *Drosophila*, it is perhaps surprising that a clear functional homolog of these proteins has not been identified in mammalian cells. However, animal cells do possess another CKI protein - called Dacapo in *Drosophila* and p27 in mammals - that helps govern Cdk activity in G1. Most importantly, p27 inhibits G1/S-Cdks (cyclin E-Cdk2), as well as the S-Cdk cyclin A-Cdk2, but has relatively little effect on the M-Cdk cyclin B-Cdkl (Russo *et al.*, 1996a). Thus, in mammals the increase in G1/S-Cdks in late G1 requires the removal of p27, which is achieved by a combination of mechanisms. First, the G1-Cdks (cyclin D-Cdk4) remove p27 from G1/S Cdks. Second, p27 is destroyed in late G1 as a result of phosphorylation by multiple protein kinases, including the G1 /S-Cdks themselves.

Other CKIs help promote G1 arrest in response to specific inhibitory signals. Far 1 in budding yeast and the INK4 proteins of mammals inhibit G1-Cdk activity when cells encounter anti-proliferative signals in the environment. The p21 protein in mammals blocks G1/S- and S-Cdks, and thus cell-cycle entry, in response to DNA damage, giving the cell time to repair the damage before starting to replicate its DNA (Harper *et al.*, 1995).

The CKIs of animal cells are grouped into two major structural families, each with a distinct mechanism of Cdk inhibition. Members of the Cip/Kip family, including Dacapo and p27, control multiple cyclin-Cdk complexes by interacting with both the cyclin and Cdk (Sherr and Roberts, 1999; Russo *et al.*, 1996a). These proteins have complex biochemical functions: they can primary block cell-cycle progression by inhibiting G1/S- and S-Cdks, as just described, but can also promote cell-cycle entry by activating G1-Cdks. In contrast, members of the INK4 family are inhibitors that display a clear specificity for the monomeric forms of Cdk4 and Cdk6 and act in part by reducing cyclin binding affinity (Russo *et al.*, 1998).

Species	Name	Alternatives	Relatives	Targets, function
S. cerevisiae	Sic1		Rum1	Inhibits S- and M-Cdks, suppresses Cdk activity in G1.
	Far1		no relatives	Inhibits G1/S-Cdks in response to mating pheromone
S. pombe	Rum1		Sic1	Inhibits S- and M-Cdks, suppresses Cdk activity in G1
D. melanogaster	Roughex/Rux		no relatives	Inhibits S- and M-Cdks, suppresses Cdk activity in G1.
	Dacapo/Dap		Cip/Kip	Inhibits G1/S-Cdks, suppresses Cdk activity in G1
X. laevis	Xic1	Kix1	Cip/Kip	Inhibits G1/S- and S-Cdks
H. sapiens	p21	Cip1/Waf1	Cip/Kip	Inhibits G1/S- and S-Cdks, activates cyclin D-Cdk4.
	p27	Kip1	Cip/Kip	Inhibits G1/S- and S-Cdks, activates cyclin D-Cdk4.
	p57	Kip2	Cip/Kip	Inhibits G1/S- and S-Cdks, activates cyclin D-Cdk4.
	$p15^{INK4b}$		INK4	Inhibits Cdk4 and Cdk6
	$p16^{INK4a}$		INK4	Inhibits Cdk4 and Cdk6
	$p18^{INK4c}$		INK4	Inhibits Cdk4 and Cdk6
	$p19^{INK4d}$		INK4	Inhibits Cdk4 and Cdk6

Table 2. Cdk-inhibitor proteins. From (Morgan, 2007).

1.3 Intrinsic disorder in Cdk-inhibitor proteins

Sic1: biological role

As previously stated, in late G1 phase, eukaryotic cells have to face a critical decision-making problem, whether to entry the mitotic cycle. This checkpoint, named 'Start' for yeast cells and 'restriction point' for mammalian cells, is a key event in the cell cycle, which commits the cell to a new round of division. The molecular bases of Start have been extensively studied over the last two decades, and many of its key molecular players and their interactions have now been identified (Morgan, 2007). In the yeast *S. cerevisiae*, the activity of Cdk1 plays a major role in cell-cycle progression (Mendenhall and Hodge, 1998). The most upstream regulator of the Start transition is the G1 cyclin Cln3 (Tyers *et al.*, 1993). When Cln3 is bound to Cdk1, the complex phosphorylates the transcriptional repressor Whi5. Whi5 is a repressor of the SCB-binding factor (SBF) and of the MCB-binding factor (MBF). Phosphorylation inactivates Whi5 and causes its exclusion from the nucleus, leading to SBF/MBF-dependent expression of a battery of genes that control the subsequent events of the cell cycle (de Bruin *et al.*, 2004). Two additional cyclins, the G1/S cyclins Cln1 and Cln2, are among the targets of SBF, and are directly involved in bud formation and spindle pole body duplication (Ubersax *et al.*, 2003). The S cyclins, Clb5 and Clb6, are among the targets of MBF and Clb5,6-Cdk1 complexes are responsible for proper timing of DNA replication (Schwob and Nasmyth, 1993). While Cln1,2-Cdk1 complexes become active as soon as the cyclins are translated, Clb5,6-Cdk1 complexes, once formed, remain inactive because of the interaction with a inhibitor protein, named Substrate/Subunit Inhibitor of Cyclin-dependent protein kinase (Sic1) (Nugroho and Mendenhall, 1994). Sic1 prevents premature DNA replication and provides correct timing for the G1-to-S transition by specifically inhibiting Clb5,6/Cdk1 complexes (Schwob *et al.*, 1994); transcription of SIC1 depends on the transcription factor Swi5 and is restricted to late anaphase (Toyn *et al.*, 1997). Yeast strains lacking Sic1 initiate DNA replication from fewer origins and show an extended S phase (Lengronne and Schwob, 2002); conversely, in strains bearing proteolysis-resistant mutants, DNA replication is delayed, indicating that the timing of Sic1 degradation determines the timing of the onset of DNA replication (Schwob *et al.*, 1994). The phenotypes of *sic1Δ* strains or Sic1 overexpressing strains hint that Sic1 contributes also to the inactivation of Clb/Cdk complexes at the exit from mitosis (Lengronne and Schwob, 2002).

The Sic1 protein is stable until late G1 when it is degraded following phosphorylation by Cln1,2/Cdk1. Its degradation requires phosphorylation of at least six of nine Cln1,2/Cdk1 consensus sites at the protein N-terminal region (Nash *et al.*, 2001). Phosphorylated Sic1 binds to Cdc4, the F-box protein of the Skp1-Cdc53-F-box (SCF) complex. SCF^{Cdc4} complex is the ubiquitin ligase that, together with the ubiquitin-conjugating enzyme Cdc34, targets Sic1 to proteosomal degradation (Nash *et al.*, 2001).
The Sic1 region required for the interaction with the Clb5,6-Cdk1 complexes, defined as kinase-inhibitory domain (KID), has been identified in the last seventy amino acids of the sequence (residues 215-284) (Hodge and Mendenhall, 1999).
Beyond the Cln1,2-Cdk1 complexes, other kinases are able to phosphorylate Sic1 and play a role modulating its function. Hog1, a stress-related kinase, stabilizes Sic1 probably by altering its binding efficacy to Cdc4 (preventing Sic1 degradation) and, thus, inhibits cell-cycle progression (Escote *et al.*, 2004). Moreover, Casein kinase 2 (CK2), an ubiquitous Ser/Thr kinase conserved in all eukaryotic cells, phosphorylates Sic1 on Ser201 and influences Sic1 ability to bind and inhibit Clb-Cdk1 complexes. In particular, it has been demonstrated that Sic1 phosphorylation by CK2 significantly increases the inhibitory activity on homologous and heterologous Clb-Cdk complexes (Barberis *et al.*, 2005b). This result hints to a regulatory mechanism that has been evolutionarily conserved between yeast and mammalian cells.
Because of the variety of the processes in which Sic1 is involved, there is an increasing interest in its characterization and in the understanding of the mechanisms underlying its activity.

Sic1: structural features

Sic1 is a 284-amino acids protein, with a theoretical molecular weight of 32,223 Da and a calculated isoelectric point of 7.89.
Structural prediction has been performed on the basis of sequence analysis (Brocca *et al.*, 2009). The consensus prediction of secondary structure, obtained with several different algorithms, hints to the existence of large portions in random-coil conformation, particularly within the N-terminal half of the molecule (Figure 4A). The prediction also includes some segments of α-helices and an overall higher content in regular secondary structure within the C-terminal moiety. Prediction of solvent accessibility indicates that very few positions have strong burial propensity, and also these are clustered in the C-terminal half of the molecule (Figure 4A). Disorder prediction by

several different algorithms indicates a strong propensity for structural disorder in Sic1 (Figure 4B) The disorder profiles for all the predictors employed, as well as the curve corresponding to the consensus score, are mostly located above the empirically defined threshold of 0.5 for the fractional score ranging between 0 and 1 (Li *et al.*, 1999), suggesting that Sic1 is a highly disordered protein, with particularly high disorder propensity in the N-terminal region.

Analogous results were obtained by comparing the Sic1 compositional profile (Vacic *et al.*, 2007) with that of the intrinsically disordered proteins from the DisProt database (Sickmeier *et al.*, 2007). Residues that are less abundant in Sic1 (or in a set of disordered proteins) than in typical ordered proteins display negative values of the fractional difference in amino acid composition, whereas enriched residue display positive values (see legend of Figure 4). Compositional profiling of Sic1 reveals that, in comparison with typical ordered proteins, Sic1 is noticeably depleted in order-promoting residues, including W, Y, C, I, L, V, M, and H, and is highly enriched in disorder-promoting residues, such as T, R, Q, S, N, P, E, and K (Figure 4C).

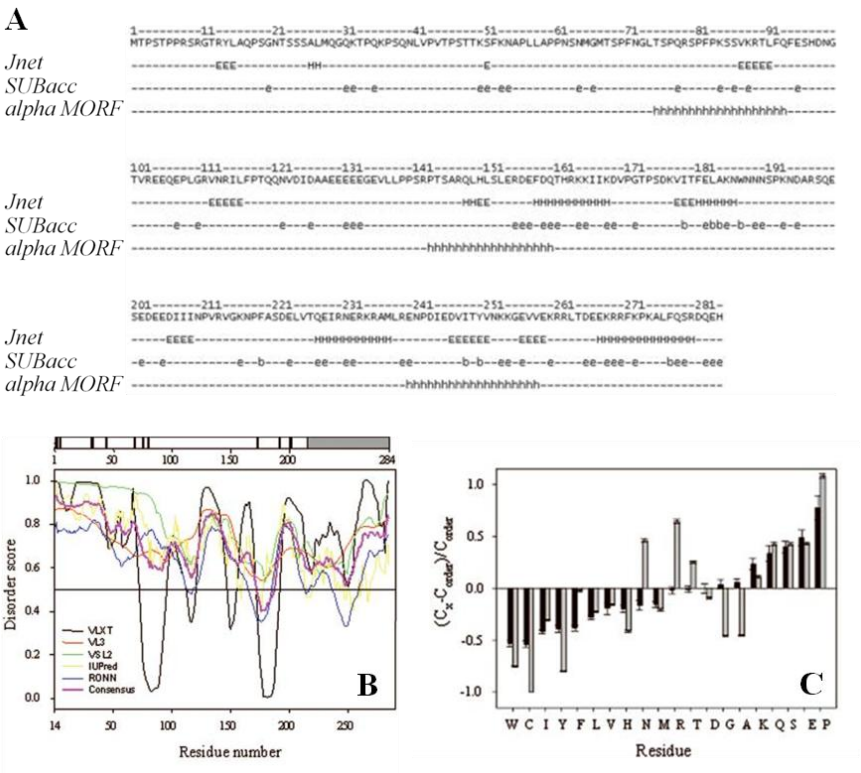


Figure 4. Sic1 disorder propensity. (A) Amino acid sequence and predicted structural features. *Jnet*, consensus secondary structure (H, helix; E, extended); *SUBacc*, high confidence accessibility prediction (e, exposed; b, buried); alpha MORF, α-helical intermolecular recognition sites predicted by α-MORF. (B) Intrinsic-disorder predictions. The plot represents prediction results by five different algorithms. The average of these five predictions is shown as a bold pink line. At the top of the panel, schematic representation of the Sic1 molecule with indication of the known phosphorylation sites (black bars) and of the inhibitory domain (shaded box). (C) Compositional profiling of Sic1. The plot shows the fractional difference in amino acid composition of Sic1 (gray bars) and of a set of intrinsically disordered proteins from the DisProt database (black bars) relative to a reference set of ordered, globular proteins. C_X: content in a given amino acid of Sic1 (or of the set of intrinsically disordered proteins); C_{order}: corresponding value in the set of ordered proteins. Negative fractional difference indicates depletion (and positive, enrichment) in the corresponding amino acid. From (Brocca *et al.*, 2009).

Sic1 conformational properties have been further investigated by far-UV CD and ESI-MS analysis, which offer complementary information on protein secondary and tertiary structure, respectively (Brocca *et al.*, 2009). The far-UV CD spectrum of Sic1 (Figure 5A) is characterized by a shoulder at 222 nm and an intense negative minimum at ~202 nm. These features indicate that Sic1 in solution is mostly disordered but contains some helical structure (~9%), in agreement with the predictions. As regards the tertiary structure, charge state distribution (CSD) analysis of nano-ESI-MS spectra, indicates that Sic1 populates a partially folded conformation (Figure 5B). The protein can be brought to a more extended conformation by the addition of denaturing agents (Figure 5C). Thus, the disordered state of Sic1 in the absence of interactors is characterized by the presence of some intrinsic tertiary and secondary structure that makes it clearly different from a completely unfolded protein. The protein displays global compactness intermediate between that of a fully unfolded and a folded globular protein and can be "denatured" by acids and organic solvents to a fully unfolded state. Thus, the properties of Sic1 in solution are reminiscent of those of molten globules or premolten globules (Uversky, 2002b).

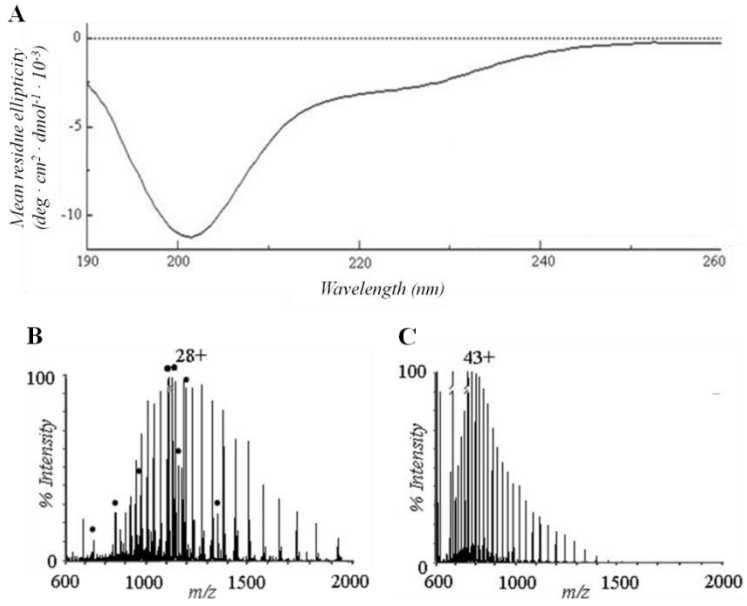


Figure 5. Sic1 secondary and tertiary structure. (A) Far-UV CD spectra of 5 μM Sic1 in 50 mM sodium phosphate pH 6.5. nano-ESI-MS Spectra of 15 μM intact Sic1 in (B) 50 mM ammonium acetate, pH 6.5; (C) 1% formic acid and 50% acetonitrile. Labels show the main charge state of each peak envelope. Peaks corresponding to Sic1 degradation products are marked by dots in the upper panel. From (Brocca *et al.*, 2009).

Limited-proteolysis experiments were used to localize intrinsically-structured regions within Sic1. The most structured moiety of the molecule is located near the C-terminus and is comprised within amino acids 178–233. Interestingly, the Sic1 proteolysis-resistant fragment overlaps with its inhibitory domain (residues 215–284). The N-terminus turns out to be the most disordered region. Structured regions or subdomains have been characterized in other IDPs and most importantly in the Sic1-related p21 and p27 (Kriwacki *et al.*, 1996; Sivakolundu *et al.*, 2005). The intrinsic structure of IDPs could play a role counteracting protein degradation *in vivo* and offering locally structured regions that may serve as nuclei for the formation of interaction surfaces.

The proteolysis-resistant C-terminal fragment of Sic1 also includes residue Ser201, the CK2 phosphorylation target. As previously described, Sic1 phosphorylation at this site has been shown to affect the G1-to-S transition and the coordination between cell size and cell cycle *in vivo*, as well as to increase Sic1 affinity for the cyclin-kinase complex *in vitro*.

On the contrary, the remarkable disorder of the Sic1 N-terminal region is likely related to the high density of phosphorylation sites, and such a correlation was observed for several other regulatory proteins (Galea *et al.*, 2008). Multiple Sic1 phosphorylation within this intrinsically disordered region mediates dynamic complex formation with the F-box protein Cdc4 (Mittag *et al.*, 2008). In particular, by an array of NMR techniques, it has been demonstrated that the high-affinity interaction between phosphorylated Sic1 and Cdc4 is mediated by multiple suboptimal threonine- or serine-phosphorylated sequences in Sic1, termed Cdc4 phosphodegrons (CPDs) (Willems et al., 2004; Mittag et al., 2008). Multiple suboptimal CPD motifs in phosphorylated Sic1 engage the core Cdc4-binding site in a dynamic equilibrium, in which multiple CPDs rapidly exchange on and off of the Cdc4-binding surface. Sites not directly bound in the core binding site at any given instant can contribute to the binding energy via a second binding site or via long-range electrostatic interactions.
Some degree of disorder can persist in IDPs even within supramolecular complexes, conferring adaptability, versatility, and reversibility to the binding reaction (Tompa and Fuxreiter, 2008). Thus, it is possible that Sic1 regions with highest disorder propensity remain locally unstructured in the bound state, thus increasing the entropy of the bound state relative to a more typical ordered complex and contributing favorably to binding (Mittag et al., 2008).

Disorder-function relationships for p21 and p27

Despite the well-documented relevance of Sic1-mediated inhibition of S-Cdk1 for cell cycle control and genomic instability, the molecular mechanism by which Sic1 inhibits S-Cdk1 activity remains obscure. Sic1 has been proposed to be a functional homologue of the mammalian Cki $p21^{Cip1}$ (Peter and Herskowitz, 1994), which in turn is characterized by significant sequence similarity with the Cki $p27^{Kip1}$ (42% identity). Therefore, it has been suggested that Sic1 is structurally and functionally related to mammalian p27, sharing a conserved inhibitory domain (Barberis *et al.*, 2005a). Moreover, evidence have been collected showing that Sic1, like p27, interacts with both the docking site and the catalytic site of the Cdk2–cyclin A complex (Barberis *et al.*, 2005a).
Among the three members of the Cip/Kip protein family (p21, p27 and p57) p27 is the most thoroughly studied. Cdk/cyclin regulation is mediated by an N-terminal KID that is conserved in p21, p27, and p57 (Figure 6A,B). The structure of p27-KID bound to Cdk2/cyclin A (Russo *et al.*, 1996a) revealed

three functional subdomains within the KID. The D1 subdomain, spanning 10 residues at the N-terminus of the KID, interacts with cyclin A through the conserved RxL cyclin binding motif. The ~ 30 residues at the C-terminus of the KID, termed the D2 subdomain, bind and inhibit Cdk2. The D1 and D2 subdomains are connected by subdomain LH, which adopts a dynamic, helical conformation. In contrast to their N-terminal KIDs, the C-terminal domains (CTDs) of Cip/Kip proteins exhibit more variability, exhibiting only partially conserved features. For instance, p27 and p57 exhibit a conserved ‘ QT ’domain within their CTDs. These domains contain phosphorylation sites, Thr187 in p27 and Thr310 in p57 (Figure 6B), that mediate interactions with the SCF Skp2 E3 ubiquitin ligase, ubiquitination, and 26S proteasomal degradation (Kamura *et al.*, 2003).

It is important to remember that Sic1 displays the same, functional regions, but the KID is located at the C-terminal region of the molecule, whereas the N-terminal moiety contains the phosphorylation sites required for the degradation.

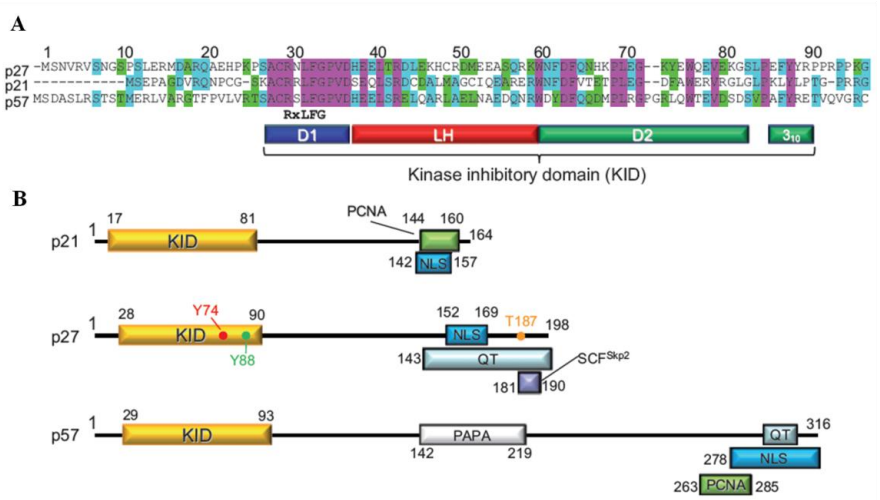


Figure 6. (A) Sequence alignment of the KID domains of the Cip/Kip proteins. The conserved subdomains are schematically represented: D1, which contains the cyclin-binding RxL motif (blue), LH (red), and D2, containing the 3_{10} helix (green). (B) Domain organization in p21, p27, and p57. The kinase binding domains are depicted in gold, the PCNA binding regions within p27 and p57 in green, and the NLSs in blue. The phosphorylation sites involved in signal transduction along p27 (Tyr74, Tyr88, and Thr187) are also labeled. From (Galea *et al.*, 2008).

Although highly disordered, as indicated by the limited resonance dispersion observed in the 2D ^{1}H- ^{15}N HSQC NMR spectrum (Lacy *et al.*, 2004), p27 retains partial secondary structure in solution. Analysis by CD indicated the

existence of a small percentage of α-helical secondary structure (Bienkiewicz *et al.*, 2002), which was subsequently attributed to the LH subdomain through NMR analysis of ^{13}C α chemical shift values (Lacy *et al.*, 2004). Furthermore, ^{1}H- ^{15}N heteronuclear NOE (hetNOE) values showed that residues within the LH and D2 subdomains exhibited partially restricted conformations. Interestingly, only D1 subdomain exhibited negative hetNOE values consistent with a high degree of flexibility. Subsequently, analysis using ^{1}H-^{1}H NOESY NMR and MD computations (Sivakolundu *et al.*, 2005) revealed that several segments within unbound p27 exhibited structural features similar to those observed in the bound state, the so-called IFSUs. This observation suggests that p27-KID associates with its binding partners through a conformational selection mechanism. This may be advantageous by reducing the entropic penalty associated with the folding upon binding. However, other regions of p27 experience binding-induced folding, giving rise to a recognition mechanism with features of both conformational selection and induced folding. In fact, isothermal titration calorimetry (ITC) and surface plasmon resonance (SPR) (Lacy *et al.*, 2004; Lacy *et al.*, 2005) were used to measure thermodynamic and kinetic parameters for p27 binding to Cdk2/cyclin A, revealing a sequential binding mechanism (Figure 7). The enthalpy of binding (ΔH) was determined using ITC. p27-KID was titrated into a solution of Cdk2, cyclin A or Cdk2−cyclin A binary complex. The heat evolved was integrated to obtain H per mole of p27-KID bound, and the binding isotherms were analyzed using a single-site binding model to determine the Gibbs free energy and entropy of binding (ΔG and ΔS, respectively). ΔG values showed that binding of p27-KID to cyclin A alone is slightly more favorable than binding to Cdk2 alone. Moreover, the values of the entropy penalty (-TΔS) observed for the two reactions were markedly different and reflected different extents of folding on binding. Indeed, D1 subdomain is relatively compact (10 residues), and therefore folding to a single conformation on binding cyclin A is associated with a relatively small entropy penalty. In contrast, D2 subdomain spans 29 residues that fold on binding to Cdk2, and, correspondingly, this reaction is associated with a large entropy penalty. Notably, D1 and D2 subdomains bound their respective targets with similar affinity, clearly revealing the thermodynamic importance of each of these interactions (Lacy *et al.*, 2004). To gain additional insights into the mechanism that governs binding of p27-KID to Cdk2−cyclin A, the kinetics of binding was studied using SPR. These studies showed that p27-KID binds cyclin A at a much higher rate than it binds Cdk2. Furthermore, binding to the Cdk2−cyclin A complex is dominated by a rate similar to that observed

for binding to cyclin A alone. It is possible that kinase binding is slow because of p27-induced remodeling of Cdk2. The kinetics of p27-KID binding to cyclin A and to Cdk2−cyclin A is biphasic, showing rapid and slow association phases. The more rapid interactions would dominate under biological conditions. The disparate kinetic behavior of p27-KID binding Cdk2 versus cyclin A strongly suggested that p27-KID binds to the cyclin A subunit of the Cdk2−cyclin A complex as the first step in a sequential binding mechanism (Lacy *et al.*, 2004; Lacy *et al.*, 2005).

First, the highly disordered and dynamic D1 subdomain rapidly binds to a conserved binding pocket on the surface of cyclin A. Next, the LH subdomain fully folds into an α -helix, which spans the gap from cyclin A to Cdk2. Three IFSUs, a β -hairpin, an intermolecular β -strand, and a 3_{10} helix, sequentially fold as p27 binds to Cdk2. Cdk2 is dramatically remodeled during p27 binding, including displacement of a β -strand by a segment of p27. It is hypothesized that the highly flexible D1 subdomain of p27 rapidly scans protein surfaces, becoming engaged on the surface of cyclin A after encountering its specific binding site. The rapid association kinetics of this interaction (Lacy *et al.*, 2004) may be due to the large capture radius of disordered p27, according to the 'fly casting' model (Shoemaker *et al.*, 2000). This fast initial step is followed by the slow component of the binding mechanism, when the longer, less flexible D2 subdomain folds onto the surface of Cdk2 and becomes positioned within the active site (Figure 7B). This sequential mechanism, which combines the induced fit and conformational selection mechanisms, seems to be conserved in p27 interactions with other Cdk/cyclin complexes. For example, kinetic analysis revealed that a truncated form of p27 lacking D1 subdomain binds to Cdk4/cyclin D at a much lower rate than does a p27 containing this subdomain. This evidence suggests that the D1 subdomain accelerates binding to this Cdk/cyclin complex, as well (Mitrea *et al.*, 2012). The sequences of the different Cdk/cyclin complexes that regulate progression of the cell division cycle are very similar, and p27 has evolved to bind the full repertoire of Cdk/cyclin complexes. Anyway, an ITC analysis of p27 binding to Cdk4/cyclin D1 revealed thermodynamic differences in comparison with binding to Cdk2/cyclin A. Although the ΔG for p27 binding to Cdk4/cyclin D1 was similar to that observed for binding to Cdk2/cyclin A, the entropic and enthalpic contributions were markedly different (Ou *et al.*, 2011). This result illustrates that, even with closely related binding partners, conserved motifs within IDPs can utilize different blends of enthalpy and entropy to drive binding.

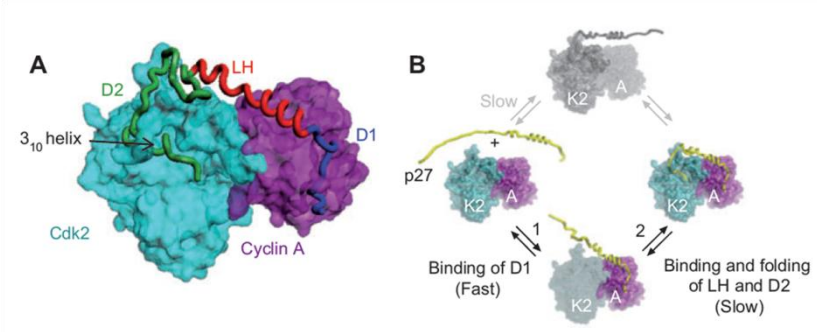


Figure 7. p27-KID binds to Cdk2/cyclin A complex in a two-step mechanism and functions as a signaling conduit. (A) Crystal structure of p27-KID in complex with Cdk2 (cyan) and cyclin A (magenta). (B) Sequential mechanism of p27-KID association with Cdk2/cyclin A. From (Galea *et al.*, 2008).

Among the three member of the Cip/Kip family of IDPs, p21 is the smallest one (164 amino acids). It binds and inhibits the full repertoire of Cdk/cyclin complexes that regulate cell division. The protein is expressed in response to genotoxic stress-induced activation of p53 and causes cell cycle arrest at the G 1/S and G2 /M checkpoints (Harper *et al.*, 1995). The sequences of the D1 and D2 subdomains of the Cip/Kip family members are highly conserved, whereas those of the LH subdomains exhibit variation (Figure 8A) (Lacy *et al.*, 2004). Structural studies using NMR demonstrated that the conformation of p21-KID bound to Cdk2/cyclin A in solution (Wang *et al.*, 2005) was similar to that of p27-KID bound to the same Cdk/cyclin (Russo *et al.*, 1996a; Lacy *et al.*, 2004). However, while some residues of the p21 LH subdomain exhibited helical resonance features, others exhibited resonance broadening due to conformational dynamics. In the p21-KID/Cdk2/cyclin A complex, modeled on the basis of the crystal structure of the corresponding complex containing p27-KID, the p21 LH subdomain was elongated, or stretched, by ~4.5 Å relative to the length of a standard 21-residues α-helix (Wang *et al.*, 2011). Furthermore, the LH subdomain of p27, in the p27-KID/Cdk2/cyclin A complex, exhibits a conformation corresponding to a stretched α-helix, elongated by about 4 Å. It has been suggested that helix stretching is responsible for the dynamic features of p21 when bound to Cdk2/cyclin A and that this 'structural adaptation' is associated with the ability to bind promiscuously to the entire Cdk/cyclin repertoire (Wang *et al.*, 2011). Experiments with p21 variants with different length LH subdomains validated this hypothesis. Furthermore, the crystal structures of Cdk4/cyclin D1 (Day *et al.*, 2009) and Cdk4/cyclin D3 (Takaki *et al.*, 2009) revealed that the LH subdomain has to contract and pivot in order for the D1

and D2 subdomains to bind conserved surfaces on the D-type cyclins and Cdk4, respectively (Figure 8). The lack of tertiary contacts between the different subdomains of the p21 an p27 KIDs reflects a flat energy landscape that enables structural adaptation while folding upon binding to Cdk/cyclin complexes.

These results also illustrate how different regions of an IDP are required to carry out different functions. In this case, D1 and D2 subdomains have evolved for specific recognition of conserved features of cyclin and Cdk surfaces, respectively, whereas the LH subdomain has evolved to adaptively span the ~ 35-Å gap between these subunits.

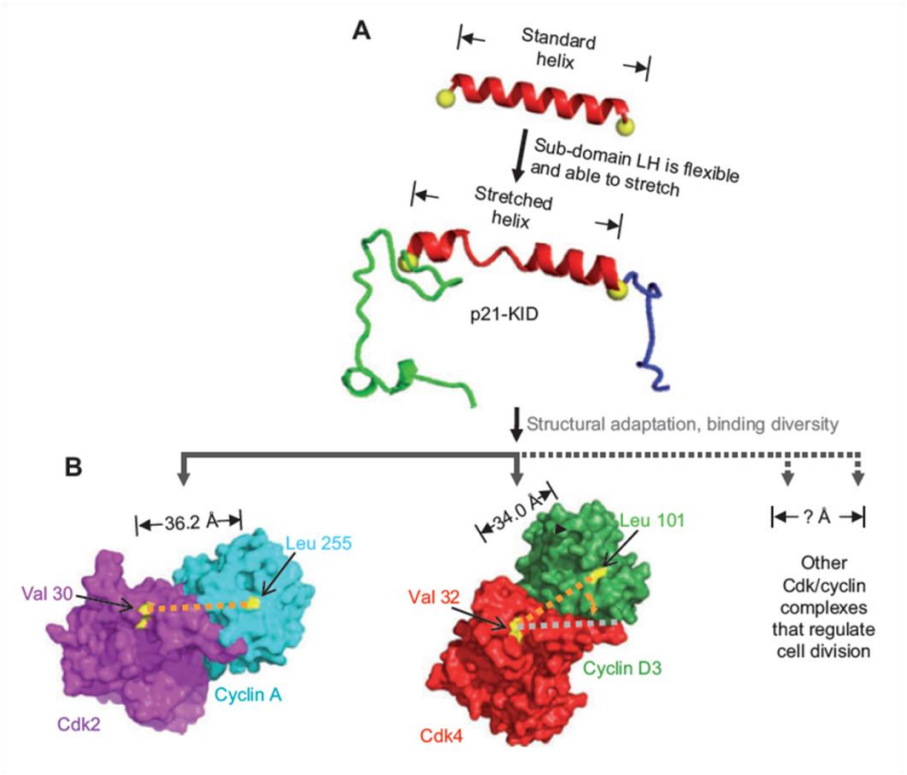


Figure 8. Structural adaptation by the LH subdomain helix in p21-mediated promiscuous binding to the Cdk/cyclin repertoire that regulates cell division. (A) Comparison of a standard α-helix, 22 residues in length, with a model of the LH subdomain of p21; structural adaptation through helix stretching accompanies binding to Cdk2/cyclin A. (B) The binding of p21 to different members of the Cdk/cyclin repertoire requires structural adaptation by the LH subdomain through helix stretching/contraction and pivoting. From (Mitrea *et al.*, 2012).

1.4 Techincal approches to study IDPs

Different experimental tools have been used in the present work to analyze protein structural features. Besides the intrinsic power of each biophysical method used, interesting information is often obtained by a synergic employment of different techniques, thanks to the complementarity of their results. For instance, the possibility to achieve both a local and a global information, or to test both average or species-specific properties of a molecular ensemble can contribute significantly to the development of a deeper knowledge of complex molecular systems.

Techniques for analysis of secondary structure

Circular dichroism (CD)

CD is widely recognized as a useful spectroscopic tool to characterize protein structure and to study protein stability. It is based on the difference in absorption by the sample of circular polarized radiations with opposite signs, clockwise and counter-clockwise. This effect occurs when the chromophores populating the solution are intrinsically chiral or are bound to a chiral centre. The sample presents a CD activity also when the chromophores are placed in an asymmetric environment as a result of the structure adopted by the molecules or of a binding reaction. In a classical CD instrument, the light coming from a xenon lamp is linearly polarized by the passage through two prisms and is focused on a piezoelectric quartz crystal subjected to an alternating electric field. In this way the crystal splits the light into its circular polarized components that are alternatively transmitted through the sample before the detection by a photomultiplier. A CD spectrum is finally obtained by plotting the dichroism signal as a function of the radiation wavelength (Kelly *et al.*, 2005). In order to remove oxygen molecules from the different compartments, a stream of nitrogen gas flows inside the instrument. This stream prevents the damage of the optical system by ozone and allows spectral measurement in the deeper far-UV region. By changing the optical path it is possible to test a very wide range of protein concentrations (from μM to mM), providing that buffer absorbance is not too high. Moreover, CD measurements can be easily performed in extreme experimental conditions, such as very acidic or basic pH, and high temperatures, making CD a fast and convenient technique for protein studies, especially in testing conformational changes (Kelly and Price, 2000).

Each region of a CD spectrum reflects prevalent absorbance of a particular kind of chromophore and thus gives specific information on the composition of the sample. For this reason, CD data are usually presented by different normalizations of a common unit, according with the fundamental absorption element in the wavelength range studied. The most adopted unit is the ellipticity (θ), defined as $tan^{-1}(b/a)$, where b and *a* are the minor and major axes of the elliptical polarized light resulting by the combination of transmitted clockwise and counter-clockwise radiations. CD measurements are usually presented in molar ellipticity, $[\theta]$, expressed in the units $deg \cdot cm^2 \cdot dmol^{-1}$. In far-UV (170-260 nm) the peptide bond is the main responsible of absorption, due to a weak but broad n→π* transition between 200 and 220 nm and a more intense π→π* transition around 190 nm (Greenfield, 1999). This comports the presentation of the data in mean residue ellipticity, $[\theta]_{mr}$, the ellipticity divided by the molar concentration, the optical path and the number of amino acid residues, all expressed in $deg \cdot cm^2 \cdot dmol^{-1}$.

The dependence of far-UV optical activity by the peptide bond geometry leads to characteristic CD spectra for the different types of regular secondary structure present in proteins (Figure 9): the α-helix motif results in an intense negative band with two peaks (at 208 and 222 nm) and in a strong positive band (at 191 - 193 nm); the β-sheets present a broad negative band at 210-225 nm and a positive peak around 190-200 nm; random coils have a negative band at 195 - 200 nm (Martin and Schilstra, 2008). The knowledge of the CD profile for each motif of secondary structure permits the estimation of the secondary structure content in a protein by different algorithms – like CONTILL, CDSSTR or SELCON – that are programmed to compare the measured spectrum with a database of protein CD spectra (Sreerama and Woody, 2000). The database is usually composed of distinct groups of globular, membrane and disordered proteins.

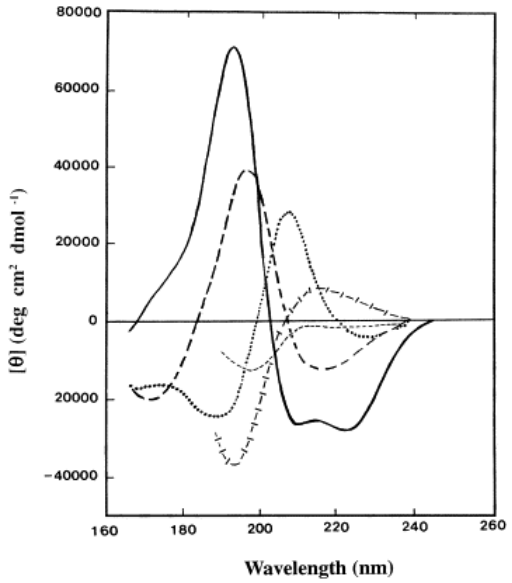


Figure 9. Far-UV CD spectra associated with various types of secondary structure. Solid line, α-helix; long dashed line, anti-parallel β-sheet; dotted line, type I β-turn; cross dashed line, extended 31-helix or poly (Pro) II helix; short dashed line, irregular structure. From (Kelly *et al.*, 2005).

Unfortunately, the characterization of protein secondary structure with these methods is not highly reliable. First of all, it must be observed that CD is a low resolution technique: only the average content of secondary structure motifs can be estimated, without any indication on the specific structure of different regions. Moreover, other chromophores distinct from the peptide bound can interfere with the signal in far-UV, as demonstrated for aromatic residues (Krittanai and Johnson, 1997). Finally, a spectrum including the deep far-UV region (170-190 nm) is necessary for an accurate secondary structure estimation, but the significant absorption of several buffers at these wavelengths makes difficult to measure much below 200 nm, unless by non-conventional instruments (Wallace, 2000).

In the near-UV (250-320 nm), protein CD signal is principally due to aromatic amino acids, with a weak contribution of disulphide bond. Phenylalanine presents a peak between 255 and 270 nm, tyrosine between 275 and 282 nm, and tryptophan between 280 and 290 nm (Strickland, 1974). The side chain of aromatic residues is particularly sensitive to the asymmetric environments characteristic of the protein tertiary structure. Nonetheless, it is not generally possible to extract any detailed information on the folded state by a near-UV CD spectrum, since several factors influence the shape and the intensity of the signal, like the number and the

relative positions of each type of aromatic amino acid, their mobility, and the nature of the environment in terms of hydrogen bonding and polarization. The near-UV CD bands from individual residues can be either positive or negative, thus a near-zero spectrum, usually typical of proteins in the unfolded state, can also be the result of mutual cancelling contributions (Griffin *et al.*, 1972). On the other hand, the near-UV CD spectrum under non-denaturing conditions represents a fingerprint of protein native tertiary structure, useful to be compared with the spectra obtained in other circumstances, like under unfolding conditions or after mutagenesis (Krell *et al.*, 1996). In particular, the presence of molten globule conformations can be easily confirmed by near-UV CD, since they are characterized by very weak signals reflecting the high mobility of aromatic side chains (Ptitsyn, 1995). Near-UV and visible are also interesting regions to monitor protein-ligand interactions. Indeed, many cofactors, like pyridoxal-5'-phosphate, heme or flavins, present no signal when free in solution, but show a band when bound to their protein partner, as a result of the acquired asymmetric environment (Price and Stevens, 1983). This feature allows one to monitor the interaction as a function of ligand concentration, or to verify the integrity of the binding site in several experimental conditions. Unfortunately, CD has no specific sensitivity to protein quaternary structure. However, protein-protein interactions can be studied by CD in some particular cases, such as when the binding comports a modification in the structure of the protein subunits, or when the oligomeric interface affects the environmental symmetry of some aromatic residues (Greenfield, 2004).

Infrared Spectroscopy (IR)

Among several optical techniques, IR has been proved to be a powerful tool to obtain information on protein secondary structure and aggregation through the analysis of the protein absorption spectrum in the mid-infrared region (wavelengths of 2.5-10 μm, corresponding to wavenumbers of 4000-1000 cm^{-1}). In particular, the amide I absorption band, due to the stretching vibration of the C=0 peptide bond from 1600 to 1700 cm^{-1}, is sensitive to the C=0 backbone environment, therefore enabling to identify the secondary structural elements of the protein (Barth, 2007). In recent years. Fourier transform infrared (FT-IR) spectroscopy has been successfully applied to IDPs for the characterization of their native state and induced folding. Indeed, as this spectroscopy allows to examine also highly scattering samples, proteins can be studied in different environmental conditions, both in solution and in solid form as protein films. This capability is very useful,

for instance, for the study of IDPs undergoing amyloid aggregation, as in the case of α-synuclein and prion peptides.

FT-IR spectroscopy is an established method to investigate protein secondary structure, stability, and aggregation through the analysis of the protein absorption spectrum. The absorption spectra of protein in water solution can be measured in spite of the high water absorption in the amide I region, using highly performing FT-IR spectrometers that, thanks to their good signal-to-noise ratio and baseline stability. allow to perform the subtraction of the solvent spectrum (Venyaminov and Prendergast, 1997). Typically, protein solutions at a concentration ranging from 1 to 15 mg/mL are required for measurements in water, while concentrations can decrease down to 0.3 mg/mL when measurements are performed in D_2O. However, the total protein amount is about 10-100 μg, since only 10-20 μL are required for a single measurement. The absorption spectra of protein solution and that of water are almost identical and dominated by the bending vibration of water around 1645 cm^{-1}. To obtain the infrared response of the protein, it is therefore necessary to perform the subtraction of the water spectrum. On the contrary, in the spectrum of a protein in D_2O solution it is possible to recognize the protein absorption before solvent subtraction since D_2O spectrum displays only a broad absorption band in the amide I and amide II region. Actually, thanks to the low absorption of D_2O in the amide I region, a good spectrum can be obtained for a protein at a concentration of about 0.5 mg/mL.

Protein solutions in water as well as in D_2O can be also measured in attenuated total reflection (ATR), a technique where the sample is placed on an ATR element having a refractive index higher than that of the sample (Goormaghtigh *et al.*, 1999). The infrared beam reaches the ATR-plate/sample-interface at an angle larger than the limit angle of total reflection. Under this condition, the beam is totally reflected by the interface and penetrates, as evanescence wave, into the sample where it can be absorbed. The optical path of the penetrating beam, which is of the order of the infrared wavelength, depends on the wavelength, on the incident angle, as well as on the refractive indices of the sample and of the ATR element. After one or several reflections, the attenuated beam is collected on the detector. This technique is very useful to keep constant, and limited to a few microns, the optical path of the sample, enabling to easily study proteins in water solutions (thanks to the sampling easiness). Also, it allows to measure proteins in the form of films, obtained by evaporation of diluted protein solutions on the ATR plate, using only a few micrograms of protein. An interesting advantage of ATR measurements is related to the possibility of

investigating the interaction of proteins with surface and synthetic membranes, and their consequent conformational changes (Goormaghtigh *et al.*, 1999).

The amide I band is a broad envelop due to the overlapping of the peptide bond absorption in the different secondary structures of the protein (Natalello *et al.*, 2005). To disclose these components, two resolution enhancement procedures can be used: the second derivative analysis of the spectra and the Fourier self-deconvolution (FSD) mehod. The second derivative analysis is a mathematical operation that can be applied to the measured spectrum after proper smoothing. This procedure enables to identify the overlapping components in the spectrum as negative bands in the second derivative. The FSD method has been proposed to resolve the different components of the protein amide I (Surewicz *et al.*, 1993). The basis of this deconvolution is that a narrowing of the bandwidth in the measured spectrum can be obtained by decreasing the decay rate of the corresponding component in the Fourier time domain. To this goal, a Fourier reverse transformation of the measured spectrum is performed, followed by a multiplication of an exponentially increasing weighting function and an appropriate apodization.

The principal absorption bands that have been widely used for protein characterization are the so-called amide I (1700-1600 cm^{-1}), amide II (1600-1500 cm^{-1}) and amide III (1400-1200 cm^{-1}), which are due to the vibrational modes of the protein backbone.

The amide I band, mainly due to the stretching vibration of the C=O peptide bond, is very sensitive to the C=O environment and, therefore, to the secondary structure of the protein. This band is the most used for secondary structure analysis. In fact, thanks to experimental FT-IR investigations and computational studies, it has been possible to assign the distinct amide I components to the protein-specific secondary structures according to their peak position (Arrondo and Goni, 1999; Barth and Zscherp, 2002). However, the absorption of the secondary structures elements can partially overlap in the spectrum, making it difficult to assign the band components. For this reason, it is crucial to validate the assignment with other FT-IR experiments or with other complementary techniques as CD.

The amide II band is due to the contribution of several backbone modes: NH in-plain bending and CN stretching, with small contributions from C=O bending, CC and NC stretching vibrations. It is also sensitive to the protein secondary structure, but its analysis is complicated by the overlapping of the different vibrational modes. Nevertheless, it has been successfully used to

study the flexibility of protein structures, since it is very sensitive to the hydrogen/deuterium exchange (Goormaghtigh *et al.*, 1999).
The amide III band is due to contribution of several vibrations: the in-phase combination of NH bending and CN stretching, with small contributions from C=O bending and CC stretching modes. It is also sensitive to the backbone structure, but for its complexity and low intensity is not extensively used. However, due to the reduced water absorption, the analysis of the amide III band may be informative for secondary structure analysis (Cai and Singh, 2004).
For a quantitative evaluation of the fraction of the different secondary structural elements of a protein, a further analysis of the spectra is required. The most common procedure is to perform a curve fitting of the amide I band as a linear combination of a set of Lorentian or Gaussian functions, each assigned to a given secondary structural element. This fitting can be performed on the measured absorption spectrum (Arrondo and Goni, 1999; Barth and Zscherp, 2002) as well as on the FSD (Goormaghtigh *et al.*, 1999). In the curve fitting of the measured spectra, the parameters of the different band components, heights, widths and positions, are adjusted iteratively in order to find a set of best fitting functions. The fractional area of each Lorentian/Gaussian function, over the component total area, represents the percentage of the corresponding secondary structure.

Techniques for analysis of tertiary structure

Nano-electrospray ionization mass spectrometry (nano-ESI-MS)

Nano-ESI-MS (Mann and Wilm, 1995) is becoming a widely employed technique in biochemistry and structural biology (Grandori *et al.*, 2009), thanks to the opportunity to preserve both covalent and non-covalent interactions while generating gas-phase ions, avoiding in this way dissociation and denaturation of biopolymers (Chowdhury and Chait, 1990). In a nano-ESI-MS experiment, ions of the analyte are transferred to the gas-phase by spraying a liquid sample from a metal-coated capillary in the presence of a strong electric field. Charged droplets, ions and ion clusters are driven through the instrument interface, where they are focused, desolvated and declustered. The process that leads from the liquid sample to gas-phase ions consists of several cycles of solvent evaporation and explosion of the droplets generated by electrospray. Indeed, the stability of a charged droplet is the result of two contrasting forces: the surface tension, that promotes droplet cohesion, and Coulomb repulsions that tend to dissociate it. During

desolvation, the latter force increases due to the increase of surface charge density. When the surface tension cannot balance electrostatic repulsions any more (the Rayleigh limit), the droplet explodes (Coulomb fission) in smaller droplets. Each daughter droplet is then below the Rayleigh limit and can undergo a new cycle of evaporation and explosion. The exact mechanism that brings to the final gas-phase analyte ion is not completely understood. In the "Charged-residue model" the evaporation-dissociation cycle continues until the droplet eventually contains only one analyte ion (Clegg and Dole, 1971). Complete solvent evaporation from such a droplet would produce a gas-phase analyte ion. On the other hand, in the "Ion evaporation model" a charged molecule is ejected from the droplet before its complete desolvatation (Iribarne and Thomson, 1976). It is thought that the charged-residue model more closely describes the process in the case of proteins (Mora *et al.*, 2000; Verkerk and Kebarle, 2005; Kebarle and Verkerk, 2009).
Numerous instrumental parameters control the ionization/desolvation process and have a critical importance in determining spray stability and spectrum quality. Most importantly, they can also affect the structure of proteins and protein complexes, and should be carefully optimized to avoid artefactual results. Among the most critical parameters are curtain gas flow rate, interface temperature and declustering potential. The curtain gas is a nitrogen stream that prevents entrance of neutral contaminants into the spectrometer and assists desolvation of ions. The collisions with gas molecules can result in protein unfolding and complex dissociation (Schneider *et al.*, 2001; Sanglier *et al.*, 2008). The instrument interface (IHT) can be heated to improve desolvation. Although it cannot be assumed that thermal equilibration is reached, by increasing the value of this parameter above room temperature it is possible to induce protein unfolding and/or complex dissociation (Invernizzi *et al.*, 2006; Natalello *et al.*, 2007). Finally, the declustering potential determines the acceleration of ions at the entrance of the instrument, where collisions with gas molecules are still possible. This parameter is important for the removal of residual solvent molecules and disaggregation of clusters. Also in this case, too high values can result in protein unfolding and dissociation (Careri *et al.*, 2004; Sanglier *et al.*, 2008), and in the alteration of the protonation states (Fligge *et al.*, 1998). As a general criterion, it is necessary to identify ranges of values in which none of these parameters becomes the limiting factor for protein stability during the electrospray process. After the desolvation and ionization steps, ions are transmitted to the mass analyzer, where they are sorted according to their mass-to-charge ratios (m/z). Quadrupoles, ion traps, cyclotron resonance traps and time-of-flight (TOF) tubes are the most

employed analyzers in MS instruments. The spectrometer used here is a hybrid quadrupole-TOF analyzer: a first quadrupole, Q0, focuses the ion beam in a second quadrupole, Q1, where ions within a given m/z range are selected and transmitted, through the last quadrupole Q2, to the TOF module. Even if not employed in this work, MS/MS measurements are possible by selecting a specific m/z value in Q1 and performing its fragmentation in Q2, which can operate as a collision cell. After the final quadrupole, the ions are pushed into the field-free TOF tube by an electrostatic plate that confers a specific kinetic energy to all of them, depending only on the applied potential and the ion charge. Then, the velocity of an ion inside the drift tube is determined only by the applied potential and its m/z value. Therefore, the ions are separated and reach the detector at different times after the pulse. To increase resolution, an electrostatic mirror (a reflectron) is employed in order to double the ion path (Benesch *et al.*, 2007). The final result is presented in a mass spectrum, that reports the (m/z) values of the ions on the x-axis and the relative intensities (ion counts) on the y-axis (Figure 10).

Proteins give rise to a bell-shaped series of peaks in the MS spectrum, called a charge-state distribution (CSD) (Kaltashov and Abzalimov, 2008). Each peak of such envelopes represents a distinct charge state and differs from the neighboring ones by one charge. The CSD of a given protein is affected by several factors, including instrumental setting and solvent properties (Mirza and Chait, 1994), but the most dramatic influence is exerted by the protein's conformational features (Chowdhury *et al.*, 1990; Kaltashov and Abzalimov, 2008). Indeed, the extent of ionization depends on the compactness of the three-dimensional structure: while folded proteins are characterized by narrow CSDs (a few peaks) at relatively high m/z values (i.e. low charge states), unfolded proteins give rise to broad CSDs (tens of peaks) shifted towards much smaller m/z values (i.e. high charge states), relative to the same protein under non-denaturing conditions.

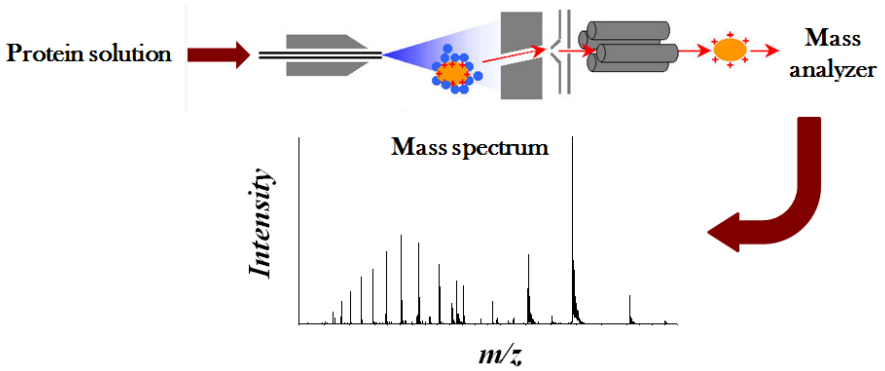


Figure 10. A representation of ion travel inside the spectrometer. Molecules are desolvated and charged during the electrospray, then focused and directed in the quadrupole. Here the ions are sorted then detected. The result is presented as plot of relative intensity (normalized ion counts) versus m/z values.

The lower average charge state of folded proteins reflects their higher structural compactness. On the other hand, the narrower CSDs of folded proteins reflect their higher conformational homogeneity (Konermann and Douglas, 1997; Grandori, 2003; Kaltashov and Mohimen, 2005).
Thanks to the fact that protein CSDs are sensitive to the global compactness of the protein structure, nano-ESI-MS also offers the possibility to detect folding intermediates (Konermann and Douglas, 1998; Kaltashov and Eyles, 2002). Compactness represents a good coordinate for folding reactions, while secondary structure or local tertiary features do not always discriminate effectively partially folded states from both native and denatured states. Multimodal distributions are generated by the coexistence of different conformational states, reflecting distinct molecular species. While the CSDs of a given protein in its folded and fully unfolded states are generally well resolved, partial overlap of CSDs can be observed in the presence of folding intermediates. Deconvolution algorithms have been developed in order to assess species distributions in conformationally complex mixtures (Dobo and Kaltashov, 2001; Borysik *et al.*, 2004). This possibility to detect distinct conformational states, free from averaging over the molecular populations, is particularly attractive. Indeed, integration of signal intensity over all the charge states of a given envelope offers a quantitative estimation of the apparent fraction of molecules in the corresponding conformational state. Unfortunately, depending on the experimental conditions, unfolded proteins can sometimes display stronger signal responses, called “relative transmission factor”, in nano-ESI-MS than folded ones, making quantitative analysis of species distributions by this method not generally applicable (Kuprowski and Konermann, 2007).

Despite these problems, nano-ESI-MS offers highly detailed qualitative information on the molecular and supramolecular species populating heterogeneous samples.

Ion mobility spectrometry (IMS)

The CSDs of individual components of protein mixtures often cover the same narrow *m/z* range so that charge states of compounds with different molecular weights may generate overlapping peaks. It can also happen that protein ions with the same mass, yet possessing different shape, share the same charge state. These ions, however, can still be distinguished by their mobility in gas phase, that is, by the so-called ion mobility spectrometry (IMS) (Cohen and Karasek, 1970). It allows to distinguish different conformers among the ions with the same *m/z* and can help reduce the spectral complexity of even complicated mixtures (Henderson *et al.*, 1999). The ions, formed by ESI or MALDI, are injected into a cell filled with drift gas (helium) and separated on the basis of their different velocity in such a buffer gas under the influence of an electric field. The velocity (v [cm/s]) of gas-phase ions through the drift cell is directly proportional to the electric field (E [V/cm]), and the proportionality constant is called mobility of the ion (K [$cm^2/(s \cdot V)$]):

$$v = KE.$$

The larger the ions (and, therefore, their collision cross section), the more hindered they will be moving through the buffer gas. Thus, the time required to traverse the drift tube is directly affected by the ion size and shape. The plot of ion intensity as a function of drift time is called arrival-time distribution (ATD). A value for the *collision cross section* can be derived from the drift time. Such a value is directly related to the shape and overall topology of the ion, yielding an information about the compactness of the observed structure. However, structural assignments based on mobility results generally require comparison with model structures obtained by molecular-dynamics simulations or molecular modeling (Kaddis and Loo, 2007).

The ion velocity v is proportional to the field strength at low electric fields (e.g., 200 V/cm). Thus, under these conditions, K is independent of E. At high electric fields (e.g., 10,000 V/cm), K starts being dependent on E in a compound-dependent manner, which is the principle of ion separation in

differential mobility spectrometry or high-field asymmetric waveform ion mobility spectrometry (Purves and Guevremont, 1999).

IMS coupled to ESI-MS provides us with a unique tool for the study of molecular conformations, allowing the separations of isomers from very complex mixtures. Nevertheless, it can be argued that gas-phase protein structures are not representative of the conformational properties in solution, since noncovalent interactions are radically affected by the presence of an aqueous solvent. Rapidly growing evidence in the literature shows, instead, that major features of protein structures and protein assemblies survive in the gas phase, at least long enough to allow species separations by IMS (Fenn and McLean, 2008; Kanu *et al.*, 2008). In agreement with experimental evidence, molecular-dynamics simulations indicate persistence of compact conformations upon desolvation of folded protein structures. These studies show that desolvated proteins tend to shrink into "inside out" structures exposing hydrophobic residues and engaging polar groups in intramolecular hydrogen bonds and salt bridges (Patriksson *et al.*, 2007). Thus, structures do change, but distinct conformational states in solution generally give rise to discernible drift times in IMS measurements.

ESI-IMS-MS has been employed in order to probe the conformation and oligomerization state of amyloidogenic proteins (Bernstein *et al.*, 2004; Baumketner *et al.*, 2006; Borysik *et al.*, 2004). The results indicate that distinct solution conformers can maintain structural differences in the gas phase, at least on the time scale of IMS measurements, allowing their discrimination based on geometric properties.

The addition of a second dimension (IMS) to MS analysis enhances tremendously the resolution power in the characterization of complex, heterogeneous mixtures.

Size exclusion chromatography (SEC)

Chromatography, a broad range of physical methods for separation and analysis of complex mixture, occupies a unique position among various analytical techniques. In chromatography, the components to be separated are generally partitioned between two phases: a stationary phase and a mobile phase. Stationary phase is typically packed into a column, whereas mobile phase usually contains dissolved sample and percolates through the stationary phase in a definite direction. The phases are chosen such that components of the sample differ in their capabilities to interact with the stationary and mobile phases. A component with higher affinity to the stationary phase will travel longer through the column than a component

with lower affinity. These differences in the mobility of sample components determine their separation as they travel through the stationary phase. SEC, also known as molecular exclusion chromatography, gel-filtration or gel-permeation chromatography (GPC), is commonly used for the separation of biomolecules, including proteins, based on their hydrodynamic dimensions. Separation in SEC is achieved via the use of the porous beads with a well-defined range of pore sizes as the stationary phase. Therefore, the separation mechanism of gel-filtration is nonadsorptive, independent of the eluent system used and very gentle. In the protein field, the most uses of SEC are separation of proteins based on their size and estimation of their molecular masses. Formally, SEC is a separation technique based on hydrodynamic radius; the hydrodynamic volume is one of the most important and fundamental structural parameters of a protein molecule. Hydrodynamic volume changes dramatically during the protein denaturation and unfolding (Uversky, 2003), and evaluation of the protein hydrodynamic dimensions (compact, extended or partially swollen) is an absolute prerequisite for an accurate classification of a protein conformation. SEC now is considered as a general separation technique where size and shape of molecules are the prime separation parameters (Porath, 1968). Therefore, the elution behavior of proteins on the SEC column is determined by their Stokes radii than by molecular masses (Uversky, 1994).

Thorough SEC-based analysis of several proteins, whose hydrodynamic dimensions in different conformational states were estimated by other hydrodynamic techniques, revealed that the SEC-determined R_s values were in good agreement with those obtained by traditional hydrodynamic methods such as viscometry, sedimentation and dynamic light scattering (Uversky, 1994). The accuracy of SEC measurements was high enough to obtain the reliable information on the hydrodynamic dimensions of a protein in different conformational states. In fact, even molten globules, whose hydrodynamic dimensions are very close to the respective values of the globular ordered proteins, were reliably discriminated from the corresponding folded proteins (Uversky, 1994). SEC was shown to represent a very useful tool to follow changes in the hydrodynamic dimensions accompanying denaturation and unfolding of globular proteins (Corbett and Roche, 1984). Importantly, it has been established that the unfolding curve retrieved for a given protein by SEC coincide with the unfolding curves measured for this protein by other techniques. This clearly indicated that the reliable R_s measurements can be done not only under the conditions preceding and following the conformational transition, but also within the transition region (Uversky, 1994; Corbett and Roche, 1984). In general, SEC

is an "inert" technique, which does not shift the equilibrium between the conformers and, therefore, can be used for a quantitative study of processes involved in protein conformational rearrangements (Corbett and Roche, 1984; Uversky, 1994).

Techniques for atomic-resolution analysis

Nuclear magnetic resonance (NMR) spectroscopy

In NMR spectroscopy radiofrequency (RF), waves in the MHz range are applied to samples immersed in a strong and homogeneous magnetic field (B_0). The magnets for this purpose are made of cryogenically cooled superconducting materials and reach fields of several tesla. In the presence of the strong external B_0 field, the nuclear spin states of any atomic nucleus which possesses either odd mass or odd atomic number (or both), attain different energies. Due to the small energy difference between the nuclear spin states, the NMR phenomenon arises from a small ($\sim 10^{-6}$) macroscopic excess nuclear spin magnetization in the sample. The NMR signal is registered as a tiny current induced in a receiver coil, due to stimulated emission from the nuclear spin system. This signal, known as the free induction decay (FID), is then Fourier transformed to resolve the frequency components that are present. The pulse-acquire scheme has paved the way to chemical structure determination. The strength of NMR is only truly unleashed when manipulating the nuclear spins between the excitation pulse and detection of the signal. Pulse sequences are a series of timed events, including delays and RF pulses, during which the nuclear spin ensamble evolves into a new state. In a general NMR experiment, the energies of interaction of the nuclear spins with internal and external electric and magnetic fields. For protein NMR, which is focused on spin ½ nuclei ^{1}H, ^{13}C and ^{15}N, electric interaction are not relevant, and the strongest interaction takes place with the external magnetic field B_0. Relevant to the application of NMR spectroscopy are the much smaller terms due to the indirect magnetic interaction of the nuclear spin with the external magnetic field through the involvement of the electrons (chemical shift), the indirect magnetic interaction of the nuclear spins with each other through the intervening electrons (scalar or J coupling), and the magnetic interactions of the nuclear spins with each other through the space (direct dipole-dipole coupling) (Mulder *et al.*, 2010).
The NMR technique allows to obtain information on the structure and

dynamic behavior of a protein (Akke, 2002). Ideally, each of the nuclei in the molecule is surrounded by a different chemical environment and therefore presents a typical chemical shift. However, proteins are macromolecules in which the number of nuclei (and therefore of peaks in the NMR spectrum) is in such a large amount that, in a mono-dimensional spectrum, the overlapping peaks make the interpretation very difficult (Abragam, 1994). Therefore, from the ^{1}H spectrum of a protein, only some general information about secondary and tertiary structure can be obtained. However, since the atomic nuclei of a protein in a compact conformation are surrounded by different chemical environments than the same nuclei in a random coil conformation, a mono-dimensional spectrum allows to identify the conformational state of an amino-acids chain. Notwithstanding the initial applications of NMR spectroscopy concerned the determination of 3D structure at atomic resolution of folded proteins in solution (Wuthrich, 1990), the technique, nowadays, is unparalleled in its ability to provide detailed structural and dynamical information on unfolded and flexible proteins (Dyson and Wright, 1998). The proton chemical shift
dispersion and line width give first-hand information on the state of proteins; poor proton chemical shift dispersion and broad lines are indicative of disorder (Chatterjee *et al.*, 2005).

For the determination of a protein structure in greater detail, multi-dimensional experiments, that allow to correlate the resonance frequencies of different nuclei, are employed; multi-dimensional spectra, in which the overlap between peaks is reduced, contain information about the connections between nuclei within a specific portion of the molecule (Wuthrich, 1986).

Homonuclear NMR spectroscopy deals with the analysis of correlations of a single isotope (generally ^{1}H). The resonance-frequency data are generated as a result of a series of mono-dimensional spectra. The different signals are then converted in a bi-dimensional spectrum, by Fourier transformation. The intensity of each peak in the spectrum is usually represented by contour lines. Experiments defined as correlation spectroscopy (COSY) and total correlation spectroscopy (TOCSY) allow to obtain information on the correlation between protons separated by up to three covalent bonds, or localized at a distance greater than three bonds from each other, respectively. In a nuclear Overhauser effect spectroscopy (NOESY) experiment, correlations between protons are identified by means of the Nuclear Overhauser effect, that is the magnetization transfer through space; the obtained 2D spectra provide information about the spatial disposition of nuclei (Wuthrich, 1990).

Heteroatom single quantum coherence (HSQC) experiments provide spectra identifying correlations between ^{1}H protons and atomic nuclei other than hydrogen. A bi-dimensional HSQC experiment contains a peak for each heteroatom-bound proton. By knowing the chemical shift of a particular proton is then possible to determine the chemical shift of the coupled heteroatom, and vice versa. The more exploited heteroatoms in this kind of experiments are ^{15}N and ^{13}C, whose low isotopic abundance, however, requires the use of labeled proteins (Gardner and Kay, 1998).

Molecular Dynamics (MD)

MD simulations are a computational method used in the theoretical study of biological molecules, which allows to calculate the behavior of a molecular system over time. MD allowed to obtain detailed information about fluctuations and conformational changes of proteins and nucleic acids, and it is now routinely used to study the structure, dynamics and thermodynamics of biological molecules and their complexes (Karplus and McCammon, 2002). Table 3 shows the timescales of the biological processes that occur in proteins.

Local motions (from 0.01 to 5 Å, from 10^{-15} to 10^{-1} s)	Atomic fluctuations Sidechain Motions Loop Motions
Rigid body motions (from 1 to 10 Å, from 10^{-9} to 1 s)	Helix Motions Domain Motions (hinge bending) Subunit motions
Large scale motions (> 5 Å, form 10^{-7} to 10^{4} s)	Helix-coil transitions Dissociation/Association Folding and Unfolding

Table 3. Timescale of protein motions. Movements amplitudes are expressed as root mean square deviation (RMSD) values. From (http://www.ch.embnet.org/MD_tutorial/pages/MD.Part1.html)

MD simulations solve Newton's equations of motion for a system of N interacting atoms:

$$m_i \frac{\partial^2 r_i}{\partial t^2} = F_i, i = 1 \dots N,$$

where m_i are the masses and r_i the positions of the atoms, and F_i the forces acting on them. The forces are the negative derivatives of a potential function V (r_1; r_2; …; r_N):

$$F_i = -\frac{\partial V}{\partial r_i}.$$

The starting conditions are the positions of the atoms, taken, for example, from a known crystal structure, and their velocities, randomly extracted from a Gaussian or a Maxwell-Boltzmann distribution at a certain temperature. Following Newton's prescription, from the initial positions, velocities and forces, it is possible to calculate the positions and velocities of the atoms at a small time interval (a time step) later. From the new positions the forces are recalculated and another step in time made. The cycle has to be repeated many times in the course of a full simulation, usually for many thousands of time steps, because a single time step is usually of the order of femtoseconds.

The system is followed for some time, taking care that the temperature and pressure remain at the required values. The coordinates as a function of time represent a trajectory of the system. After initial changes, the system will usually reach an equilibrium state. By averaging over an equilibrium trajectory, many macroscopic properties can be extracted from the simulation (Karplus and McCammon, 2002).

The potential energy function of a system is a function of all the atomic positions, and can be considered as a sum of bonded and nonbonded terms. The bonded potential terms involve 2-, 3-, and 4-body interactions of covalently bonded atoms, and describe bond stretching, angle bending and torsion potentials. The nonbonded potential terms involve the interactions between atoms not bound to each other or separated by more than three covalent bonds, both van der Waals and electrostatic interactions. All classical simulation methods rely on more or less empirical approximations called force fields to calculate interactions and evaluate the potential energy of the system as a function of the atomic coordinates. A force field consists of both the set of equations used to calculate the potential energy and forces from particle coordinates, as well as a collection of parameters used in the equations (MacKerell *et al.*, 1998).

The solvent, usually water, exerts a major influence on structure, dynamics and thermodynamics of biological molecules, both locally and globally. One of the most important effects of the solvent is mitigation of electrostatic interactions. It is important to include the effects of the solvent in an MD simulation. This can be done at different levels: in the simplest case it is possible to include a dielectric constant in the electrostatic term of the potential function. In this implicit solvent consideration, the water molecules

are not included in the simulation, but their effect is mimicked using the dielectric constant. If, instead, the water molecules are explicitly included in the simulation, the potential evaluation is more accurate (Vorobjev *et al.*, 1998).

Essential Dynamics (ED)

Functional proteins are mechanical objects in which only some types of internal motions are allowed to carry out their biological functions. The internal motions may be related to the binding to a substrate or a coenzyme, to the adaptation to different environments or to the transmission of conformational adjustments to influence the reactivity in a remote site of the structure (Amadei *et al.*, 1993). These internal motions can be very feeble and involve complex relations at atomic level, their nature, however, is inherent in the total structure and in the interactions within the molecule. Identifying these motions by starting from the molecular structure would allow to reduce the complex dynamics of a protein to its essential degrees of freedom.

The conformational freedom is related to the potential energy: theoretically, the distribution of conformations at equilibrium can be calculated from the potential. However, proteins are systems too complex for doing so, because of their "high dimensionality" and because they have generally potential energy surface with many minima (Hess, 2000).

ED, that is the principal component analysis (PCA) applied to MD simulations, is a powerful tool for evaluate both the dynamic properties of a system and the efficacy of conformational-space sampling. PCA is a mathematical technique that allows to analyze sets of multidimensional data, and to define a new coordinate system, in which there is no covariance associated with each pair of coordinates. In this sense, the new coordinates can be defined uncorrelated and ordered according to the variance content in the data; this allows to reduce the dimensionality of the analyzing space, simply by discarding the coordinates with less variance and by focusing on those that describe the greater and wider fluctuations (Hess, 2000).

It is possible to separate the conformational space into two subspaces (Amadei *et al.*, 1993): an *essential subspace*, containing only a few degrees of freedom, in which anharmonic motion occurs that comprises most of the positional fluctuations; a *constrained subspace*, the remaining space in which the motion has a narrow Gaussian distribution, due to small, high-frequency fluctuations, and which can be considered as physically constrained.

Since its introduction (Amadei *et al.*, 1993), ED has proven to be a useful technique to identify protein regions in which most of the functional motions are localized, and it has been successfully applied to describe the dynamic features of many proteins.
Molecular dynamics is a suitable tool to identify structure-function relationships in a protein: from the calculated trajectories, structural and dynamical properties can be extracted. A limitation of this method is that the simulations are limited in time and therefore the sampled conformations are only a part of all those that a protein can assume. To correlate these data with the function of the protein, therefore, is necessary to ensure the sampling efficiency. By performing several replicas, that is simulations of equal or different duration starting from the same structure, and by employing the ED technique to analyze the conformational space sampled during the simulations, it is possible to verify the reliability of the trajectories and, therefore, of the related results.

Free energy landscape (FEL)

Proteins are very complex systems and the FEL that describes the native state actually contains many points of minimum, corresponding to slightly different conformations, defined conformational substates (Henzler-Wildman and Kern, 2007). MD simulations, providing the information at atomic level about the motions of proteins, allow the representation of the FEL (Lei *et al.*, 2007). However, since it is not possible to represent the FEL as a function of 3N-6 coordinates, which correspond to the total degrees of freedom of a system with N atoms, is fundamental choosing a set of reaction coordinates that allows to distinguish between different conformational substates. ED is a powerful tool to derive significant reaction coordinates. Once the reaction coordinates have been chosen, the FEL can be calculated from the probability density function. This is defined as the function describing the relative likelihood for a random variable to take on a given value. To achieve a two-dimensional representation of the FEL, the probability of finding the system in a particular state characterized by a value q_α of some variables of interest, the reaction coordinate, can be defined as proportional to $e^{-G\alpha/kT}$, where $G\alpha$ is the free energy of the state. The FEL can be obtained from:

$$Ga = -kTln\left[\frac{P(q_a)}{P_{max}(q)}\right],$$

where k is the Boltzmann constant, T is the temperature of simulation, $P(q_\alpha)$ is an estimate of the probability density function obtained from a histogram of the MD data and $P_{max}(q)$ is the probability of the most probable state. Considering two different reaction coordinates q_i and q_j, the two-dimensional FEL can be obtained from the joint probability distributions $P(q_i,q_j)$ of the system (Papaleo *et al.*, 2009).

2. EXPERIMENTAL WORK

In this section different experimental works are presented.

2.1 Charge-surface correlation in electrospray ionization of folded and unfolded proteins.
Testa L, Brocca S, Grandori R.
Anal Chem. 2011 Sep 1;83(17):6459-63.

2.2 Defining structural domains of an intrinsically disordered protein: Sic1, the cyclin-dependent kinase inhibitor of Saccharomyces cerevisiae.
Brocca S, Testa L, Samalikova M, Grandori R, Lotti M.
Mol Biotechnol. 2011 Jan;47(1):34-42.

2.3 Electrospray ionization-mass spectrometry conformational analysis of isolated domains of an intrinsically disordered protein.
Testa L, Brocca S, Samalikova M, Santambrogio C, Alberghina L, Grandori R.
Biotechnol J. 2011 Jan;6(1):96-100.

2.4 Compaction properties of an intrinsically disordered protein: Sic1 and its kinase-inhibitor domain.
Brocca S, Testa L, Sobott F, Samalikova M, Natalello A, Papaleo E, Lotti M, De Gioia L, Doglia SM, Alberghina L, Grandori R.
Biophys J. 2011 May 4;100(9):2243-52.

2.5 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, Grandori R.
Front Physiol. 2012;3:435.

2.6 Heterologous expression, purification and characterization of Sic1-physiological interactors, Cdk1 and Clb5.
Testa L, Grandori R, Brocca S
Preliminary results

As widely discussed in the introduction, many naturally occurring proteins have been shown to lack rigid three-dimensional structure, existing instead as dynamic ensembles of inter-converting conformations. In isolation, these proteins exhibit a highly dynamic structure that is resembling more the denatured rather than native state of "normal" globular proteins. Intrinsically disordered proteins (IDPs) have attracted a great deal of interest since it became clear that their lack of structural specificity is of physiological importance. Indeed, structural disorder characterizes a broad class of regulatory proteins, which all share the feature to bind multiple interactors. Intrinsic disorder is a common property of many protein types especially among the proteins involved in signal transduction and transcription regulation. Structural characterization of proteins in disordered conformation is technically difficult, but it is important to better understand folding transitions to ordered states. A promising approach to this technically challenging problem is offered by a multiparametric analysis, which employs a set of complementary biophysical methods sensitive to distinct structural features.

One of these techniques, electrospray ionization-mass spectrometry (ESI-MS), has developed into a central tool of biochemistry and structural biology, and is widely used for protein studies. However, the mechanism determining protein ionization during electrospray is not well understood yet. It has been shown that the extent of protein ionization under nondenaturing conditions correlates well with the solvent-accessible surface area (SASA) of the tridimensional structure, for either folded monomers or multimeric complexes. The work reported in section 2.1 represents a study to test if this charge-surface correlation holds for unfolded proteins as well. The SASAs for a set of proteins in the unfolded state were calculated on the basis of their amino acid sequences, and the results were used to analyze the available ESI data, either extracted from the literature or produced in our laboratory, in terms of protein surface. The charge-to-surface relation seemed to hold for unfolded proteins, as well as it was previously demonstrated for folded ones. Conversely, when the extent of ionization is analyzed as a function of protein mass, folded and unfolded proteins yield very different profiles. IDPs have proven to be very useful in this study. Indeed, since the nature of the solvent can affect the protein charge-state distributions (CSDs), it was important to rule out that the comparison between folded and unfolded proteins might have been biased by solvent effects generated by the different conditions employed for electrospray. IDPs, being unfolded under nondenaturing conditions, offered the proper control to this regard. Data points for the high-charge components of the

multimodal charge-state distributions (CSDs) of IDPs displayed very similar dependence of average charge on protein mass as chemically denatured proteins, indicating negligible solvent effects. The general charge-surface correlation identified by this work can help rationalize conformational effects in protein ESI-MS, since protein unfolding causes a dramatic increase in the SASA and implies higher protein ionization. Moreover, a novel tool for extracting structural information from ESI-MS data has been developed. Indeed, the technique can be employed to estimate the SASA of a protein, as it will be shown in section 2.5, or to formulate hypotheses on the fold state of unknown proteins.

In section 2.2 the results concerning the identification of Sic1 as a modular IDP are presented. In the yeast *S. cerevisiae*, Sic1 controls the timing of entrance in S phase during the cell cycle by inhibiting the Cdk1-Clb5/6 complexes, whose activity is required for the G1-to-S transition. A previous, structural characterization of full-length Sic1 by limited proteolysis and mass spectrometry had helped to localize its most compact domain in the C-terminal region. Since the fragments showing the highest relative resistance to proteases started at position 172 and 176, fragments 172–284 ($Sic1^{\Delta171}$) and 176–284 ($Sic1^{\Delta175}$) were initially produced. Fragment 215–284 ($Sic1^{\Delta214}$), corresponding exactly to the previously identified functional kinase-inhibitor domain (KID), was also produced. The recombinant, histidine-tagged proteins were expressed in *E. coli* cells. These protein fragments spanning the C-terminal region of Sic1 were all expressed at high amount; further mass spectrometric investigation, anyway, indicated that $Sic1^{\Delta171}$ and $Sic1^{\Delta175}$ were unstable at their N-terminus, suggesting propensity to generate a shorter Sic1 C-terminal fragment starting from residue 187. The Sic1 fragment 187–284 ($Sic1^{\Delta186}$), subsequently produced, displayed no evidence of degradation. Therefore, $Sic1^{\Delta186}$ and $Sic1^{\Delta214}$ were selected as fragments representing the C-terminal domain, according, respectively, to a more structural or functional definition of protein domain, as derived from limited-proteolysis and *in vivo* complementation studies. Meanwhile, a N-terminal fragment, containing residues 1–186 ($Sic1^{1–186}$) was designed as the complementary fragment to $Sic1^{\Delta186}$. Fragments $Sic1^{1–186}$ and $Sic1^{\Delta186}$ were analyzed by different biophysical techniques. Circular dichroism (CD) measurements and nuclear magnetic resonance (NMR) spectra confirmed their disordered nature but reveal minor structural differences that may reflect their distinct functional roles. Nano-ESI-MS was employed, too, in order to perform conformational analyses, and the work presenting the results is reported in section 2.3. Bimodal CSDs were observed, characterized by a narrow, low-charged component and a broad,

high-charged one. Such CSDs indicated the coexistence of distinct conformers in solution for both the fragments. The narrow, low-charged component was present when the proteins were analyzed under nondenaturing conditions, suggesting that both fragments transiently populate ordered, compact states. Interestingly, the putative compact form turned out to represent only 2% of the population for the N-terminal domain, while it was estimated around 29% for $Sic1^{\Delta 186}$. Conformational transitions, induced by the addition of denaturants, provided additional evidence that the putative compact forms can be "denatured" by acids and organic solvents into more extended conformations. The highly charged components in the distributions observed under nondenaturing conditions indicated the presence of disordered species. Addition of denaturants revealed that, as regards $Sic1^{\Delta 186}$, this component represents already an almost completely unfolded state, whereas the maximally extended conformation of $Sic1^{1-186}$, represented by a quite heterogeneous molecular population, could be reached only by chemical denaturation. Therefore, ESI-MS results, too, confirmed that the two domains exhibit slightly different conformational properties.

Sections 2.4 and 2.5 report works in which a detailed description of the compact conformation of the Sic1 inhibitory domain in the absence of interactors was achieved. As already stated, the structural characterization of IDPs in their free state is essential to the interpretation of their biochemical properties. Due to their dynamic and heterogeneous conformation, anyway, such a characterization is technically challenging, so $Sic1^{\Delta 214}$ fragment has been analyzed by an array of biophysical techniques (section 2.4) and by all-atom molecular dynamics (MD) simulations (section 2.5). This approach allowed to describe the conformational properties of Sic1 KID in its free state, both in terms of secondary and tertiary structure. Secondary structure analysis, performed by Fourier-transformed infrared spectroscopy (FT-IR), highlights the disordered nature of $Sic1^{\Delta 214}$, with a predominance of random coil and a certain intrinsic propensity to form ordered secondary structure elements, primary α-helical structures and 3_{10}-helices. Then, the hydrodynamic behavior of the protein was characterized and its hydrodynamic radius (R_h) determined. This result, together with the evidence that R_h values of folded, globular proteins and those of chemically denatured proteins display a well-characterized dependence on the number of residues, was employed to define a new parameter, the compaction index (CI). The CI allows to compare the degree of compaction among proteins with different lengths. In particular, CI was used to compare the degree of compaction of $Sic1^{\Delta 214}$ and the full-length Sic1, and the results indicated that the full-length protein is endowed with higher overall compaction than the

fragment. Interestingly, such a conclusion implies that either long-range coupling between the N- and C- termini or indirect effects of chain length play a role in the compaction of the full-length protein. Finer characterization of tertiary structural features of $Sic1^{\Delta 214}$ was achieved by ESI-MS. As shown for $Sic1^{\Delta 186}$ in section 2.3, the spectra of $Sic1^{\Delta 214}$ under nondenaturing conditions were characterized by sharply bimodal distributions, with a predominant form and a minor component, reflecting an extended and a compact conformer, respectively. Again, conformational transitions provided further evidence that the minor component in the CSD represents a metastable compact state of the molecule in solution. The existence of compact conformations was then tested by electrospray-ionization ion-mobility mass-spectrometry (ESI-IM-MS) measurements. By adding a second dimension to conventional ion sorting by ESI-MS, IM leads to separation of the ions on the basis of their structural compactness. The results provided direct evidence of the existence of compact conformations of $Sic1^{\Delta 214}$, clearly distinguishable from the extended form of the protein. In order to achieve a model of the compact conformations detected by the array of biophysical techniques, the conformational ensemble of Sic1 KID was investigated by integrating all-atom explicit solvent MD simulations with experimental data. Multiple, 50–100 ns, independent MD simulations were carried out starting from extended models generated by satisfying constraints on the secondary structure. More than 500 ns of MD trajectories were collected, and the results were analyzed with reference to secondary structure content, SASA and intramolecular interactions. Moreover, a statistical analysis on the trajectories allowed to perform free energy landscape (FEL) calculations and to identify distinct conformational substates in the simulated ensemble. The MD data were analyzed to identify putative intrinsically folded structural units (IFSUs), that is regions that were characterized by at least transient secondary structure. Some of the average structures identified in the simulated ensemble provided good agreement with the FT-IR data discussed in section 2.4. These models displayed a finger-like structure, composed by α-helices of different length. The tertiary structure properties of the distinct conformations populated by Sic1 KID were characterized by calculating their SASA values. Importantly, the results could be compared with estimates from ESI-MS data, derived by the charge-surface correlation described in section 2.1. The computational values compare quite well with the experimental data concerning the more compact form, an expected result because classical MD simulations are able to capture conformational properties of collapsed forms but leaves extended states quite unexplored. To investigate the driving force that likely promotes

structural compaction in Sic1 KID, the different types of non-covalent interactions were examined for each conformational substate. Salt-bridge, aromatic, hydrophobic, and amino-aromatic interactions were analyzed for persistence over the simulation time. This analysis identified salt bridges as the major factor stabilizing compact structures, with a large number of basic and acidic residues involved in multiple interactions. Conversely, hydrophobic-interactions were characterized by very low persistence and, again, this result was validated by ESI-MS experiments. Indeed, unlike the denaturing effects of acids shown in section 2.4, the addition of acetonitrile or methanol up to 50% did not significantly affect the CSD, leaving clearly preserved the compact state. In conclusion, these results indicated that Sic1 KID, in its unbound state, is likely to be characterized by intrinsic secondary and tertiary structure, and pointed out a predominant role of electrostatic interactions in promoting protein compaction.

Since protein interactions are crucial for protein function, and many functions of IDPs involve specific protein-protein interactions, a full understanding of binding on thermodynamic and structural levels requires detailed information both on the free states of the interacting partners and on the complex. No structural data are available for the Sic1-Cdk1-Clb5/6 ternary complex, so the molecular mechanism by which Sic1 inhibits S-Cdk1 activity remains unclear. A model have been proposed, according to which Sic1 acts similarly to p27, an IDP involved in the regulation of the G1-to-S transition in mammalian cell cycle. Nevertheless, a better understanding of the mechanism of interaction is required, in order either to rationalize physiological, *in vivo* results, and to achieve a structural and thermodynamic description of binding. To do so, the physiological partners of Sic1 have to be purified and characterized. In section 2.6, preliminary results about the heterologous expression and purification of Cdk1 and Clb5 are presented. Both interactors have been expressed as histidine-tagged proteins in *E. coli* cells. A soluble form of Cdk1 could be extracted by treating the cells with a gentle lysis buffer, and purified by affinity chromatography. CD analysis indicated that the protein acquires the correct fold to carry out its biological function. Conversely, full-length Clb5 underwent proteolytic degradation when expressed in a recombinant way, probably because of the high degree of disorder in the N-terminal region of the amino-acids sequence. The deletion of the first 156 residues allowed the expression of the protein as inclusion bodies (IBs). A purification procedure was then developed, so that IBs were solubilized and refolded by on-column removal of denaturing conditions. The preliminary results presented in section 2.6 represent promising starting points for the production of a

functional Cdk1-Clb5 complex and so, ultimately, for the complete description of Sic1 binding mechanism.

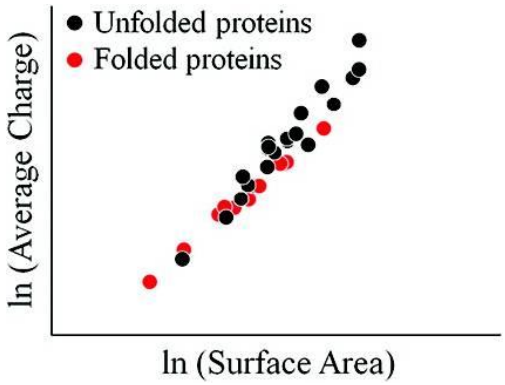
Charge-Surface Correlation in Electrospray Ionization of Folded and Unfolded Proteins

Analytical Chemistry. 2011 Sep 1;83(17):6459-63

Lorenzo Testa, Stefania Brocca, and Rita Grandori*

Department of Biotechnology and Biosciences, University of Milano-Bicocca, Piazza della Scienza 2, 20126 Milan, Italy

ABSTRACT



Electrospray-ionization mass spectrometry (ESI-MS) is widely used for protein studies. It has been shown that the extent of protein ionization under nondenaturing conditions correlates well with the solvent-accessible surface area of the tridimensional structure, for either folded monomers or multimeric complexes. The goal of this study was to test whether this relation holds for unfolded proteins as well. In order to overcome the paucity of structural data, the server ProtSA was used to model the conformational ensembles of proteins in the unfolded state and generate estimates of the average solvent accessibility. The results are analyzed along with literature data or original measurements by ESI-MS. It is found that the charge-to-surface relation holds for proteins in the unfolded state, free from solvent effects. A double-log plot is derived, in close agreement with published data for folded proteins. These results suggest that the solvent-accessible surface area is a key factor determining the extent of protein ionization by electrospray, independent of the conformational state. This conclusion helps rationalizing conformational effects in protein ESI-MS. The here reported relation can be used to predict the average solvent accessibility and, hence, the state of folding of unknown proteins from ESI-MS data.

Electrospray ionization-mass spectrometry (ESI-MS) has developed into a central tool of biochemistry and structural biology.[1,2] However, the mechanism determining protein ionization during electrospray is not well understood yet.[3] Identifying the limiting factor in protein ionization by electrospray is essential to confidently interpret MS data for protein structural investigation. For folded, globular proteins, sprayed under nondenaturing conditions, there is a well-known charge-to-mass correlation.[4-7] The experimentally observed net charge increases with protein size approximately as $R^{1.5}$, where R is the radius of the sphere that approximates the tridimensional structure. For folded, globular monomers, a correlation with mass obviously implies a correlation with surface. However, it has been pointed out that, considering shapes with altered mass-to-surface ratios, for instance elongated structures or hollow protein complexes, it is the relation to surface to hold, rather than the relation to mass.[7-9] Unfolded proteins respond apparently differently from folded proteins to the electrospray process. The same protein acquires many more charges in the denatured state, rather than in a native-like conformation.[8,10] The goal of this study was to test whether the relation to surface that describes ionization of folded proteins could be extended to unfolded proteins, indicating a general dependence on the solvent-accessible surface area of the protein structure.

EXPERIMENTAL SECTION

The protein BLG (see Table 1 for protein abbreviations) was purchased from Sigma Aldrich (St. Louis, MO) and rinsed on Amicon-Ultra filter devices (Millipore, Carrigtwohill, Ireland) with a 10000 Da cutoff. The intrinsically disordered proteins (IDPs) NTAIL and PNT (see Table 1 for protein abbreviations) were produced as recombinant proteins as previously described.[11-13] Nano-ESI-MS measurements were performed as already described.[14,15] Samples were infused at room temperature in 50 mM ammonium acetate, pH 6.5.
The average charge state was calculated from mass spectra as

$$z_{av} = \frac{\sum_{i=A}^{B} z_i I_i}{I_{tot}}$$

where A and B are the lowest and the highest charge states of a distribution, *z* is the charge state, and *I* is signal intensity measured as peak height. Protein mass was calculated from the amino acid sequence, by means of the Expasy tool Compute pI/MW (http://expasy.org/tools/pi_tool.html). The contribution of heme was not included in the calculation for hemoproteins. Solvent-accessible surface area was calculated by the ProtSA server (http://webapps.bifi.es/protsa),[16] giving as input the PDB code for folded globular proteins or the amino acid sequences for unfolded proteins.

Table 1. Proteins Used in This Work [a]

protein name [b]	molecular weight	fold type	ESI conditions [c]	ref (PDB)
Abeta (1–40)	4329.86	IDP	D	17
IB5	6923.74	IDP	D	18
ACP_L46W	8604.43	IDP	N-D	19
$Sic1^{\Delta 214}$	9293.38	IDP	N-D	14
Sml1	11834.19	IDP	D	20
prothymosin-α	12073.85	IDP	N-D	21
$Sic1^{\Delta 186}$	12676.00	IDP	N-D	22
α-synuclein	14460.16	IDP	N-D	8
NTAIL	14633.00	IDP	N-D	this work
$Sic1^{1-186}$	21593.13	IDP	D	22
PNT	24821.00	IDP	N-D	this work
xXPA	30922.45	IDP	D	23
$Sic1^{full-length}$	33102.88	IDP	D	24
ubiquitin	8564.84	globular	D	25
cytochrome c	11701.55	globular	D	19,25,26
β2-microglobulin	11862.36	globular	D*	15
lysozyme	16238.65	globular	D*	26
myoglobin	16951.48	globular	D	27
BLG	18265.00	globular	D*	this work
BGL	33091.87	globular	D	28
insulin	5721.57	globular	N-D	7 (1APH)
ubiquitin	8564.84	globular	N-D	7 (1UBQ)
CRABP1	15591.60	globular	N-D	7 (1CBI)
myoglobin	16951.48	globular	N-D	7 (1WLA)
BLG	18367.30	globular	N-D	7 (1BSY)
H-chain of ferritin	21225.64	globular	N-D	7 (2FHA)
α1-chymotrypsinogen	25666.12	globular	N-D	7 (1EX3)
carbonic anhydrase I	28753.04	globular	N-D	7 (2CAB)
holotransferrin, N-lobe	36360.40	globular	N-D	7 (1A8E)
BLG dimer	36600.51	globular	N-D	7 (1BEB)
apotransferrin, N-lobe	37167.27	globular	N-D	7 (1BTJ)
hemoglobin	61933.10	globular	N-D	7 (1A3N)
ACP	*8531.37*	*IDP*	*N-D*	19
X-Lin-DnaB-N	*13690.28*	*globular*	*N-D*	29
α-lactalbumin	*14223.17*	*globular*	*N-D**	29
α-synuclein	*14460.16*	*IDP*	*N-D*	30
lysozyme	*16238.65*	*globular*	*D**	26
RNase A	*16460.88*	*globular*	*D**	29
calmodulin	*16706.39*	*semidisordered*	*N-D*	31
myoglobin	*16951.48*	*globular*	*D*	32
calbindin	*29994.07*	*semidisordered*	*N-D*	33

[a]Data from measurements in the negative-ion mode are reported in italics. [b]ACP, acyl carrier protein; Sic1, substrate/subunit inhibitor of cyclin-dependent protein kinase; $Sic1^{\Delta 186}$, Sic1 fragment from amino acid 187 to 284; $Sic1^{\Delta 214}$, Sic1 fragment from amino acid 215 to 284; $Sic1^{1-186}$, Sic1 fragment from amino acid 1 to 186; Sml1, suppressor of Mec1 lethality; NTAIL, C-terminal domain of the measles virus nucleoprotein N; PNT, N-terminal domain of measles virus phosphoprotein P; xXPA, Xenopus xeroderma pigmentosum group A protein; BGL, Burkholderia glumae lipase; BLG, bovine lactoglobulin; CRABP1, cellular retinoic acid binding protein 1; X-Lin-DnaB-N, linear version of cyclized DnaB helicase from Escherichia coli.[c] D, denaturing conditions; N-D, nondenaturing conditions; the asterisk indicates that the protein was reduced in the presence of DTT.

The default values of 1.4 Å for solvent radius and 2000 as the number of conformations to be generated were used for each run. The sequence used for the calculations included the histidine tag when required. The data of Figure 1 were fitted by a power function $y = ax^b$, while the data of Figures 2 and 3 were subjected to linear fitting. Fitting analyses were performed by the software Origin 8 (Originlab, Northampton, MA).

RESULTS AND DISCUSSION

Folded, globular proteins sprayed under nondenaturing conditions are compared here to proteins denatured by organic solvents. The proteins used in this study are listed in Table 1. The data are either extracted from the literature or produced in our laboratory. Figure 1 shows that, when the extent of ionization is analyzed as a function of protein mass, folded and unfolded proteins yield very different profiles. Both sets of data can be fit by power functions, similar to what was previously reported for folded proteins.[4-7] However, the power coefficients derived by the fitting differ substantially (0.5753 for folded and 0.7368 for unfolded), reflecting the higher propensity to acquire charges of proteins in the denatured state. Since the nature of the solvent can also affect the protein charge-state distributions (CSDs), it is important to rule out that the comparison between folded and unfolded proteins might be biased by solvent effects generated by the different conditions employed for electrospray. Intrinsically disordered proteins (IDPs) offer a good control to this regard, being unfolded under nondenaturing conditions. Coexistence of highly disordered and collapsed conformers is typically reflected by bimodal or multimodal CSDs. To our purpose, always and only the high-charge component was used to calculate the reported values of average charge. Data points for such highly disordered components of IDPs in water or ammonium acetate are overlaid to the plot in Figure 1. These points display very similar dependence of average charge on protein mass as chemically denatured proteins, indicating that solvent effects are negligible in the context of the present discussion.

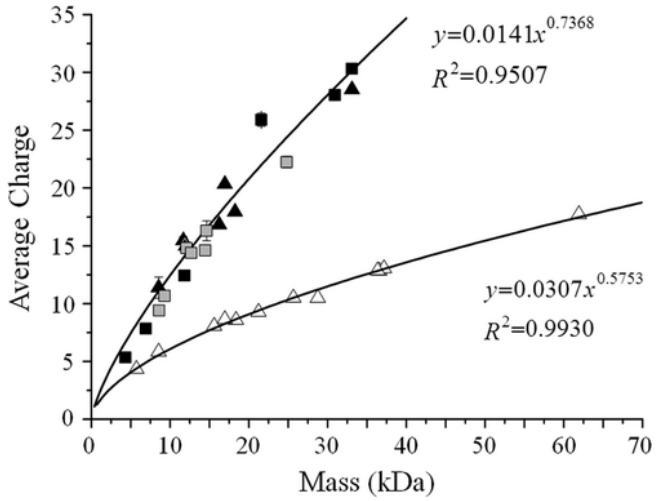


Figure 1. Charge-to-mass correlation of folded globular proteins and unfolded proteins (IDPs and chemically denatured proteins). The experimental average charge in the positive-ion mode is plotted against the calculated protein mass. Black squares, IDPs under denaturing conditions. Black triangles, globular proteins under denaturing conditions. Gray squares, IDPs under nondenaturing conditions (in the case of multimodal distributions, only the high-charge component was considered). The latter points are overlaid to the plot as a control for solvent effects and were not used to calculate the linear regression. Open triangles, globular proteins under nondenaturing conditions. Error bars are given only for the data obtained in our laboratory and represent the standard deviation from three independent experiments. The power functions derived by data fitting, and the R-squared values are indicated on the figure.

The next step is to analyze the available ESI data in terms of protein surface. Unfortunately, calculations of the solvent-accessible surface area for proteins in the denatured state are impaired by the paucity of structural models. To overcome this problem, surface values were estimated by the ProtSA program.[16] This web application generates conformational ensembles for proteins in the unfolded state on the basis of the amino acid sequence and calculates the average solvent accessibility of such an ensemble. For proteins with known structure, the surface area can be calculated using PDB files as input. The ProtSA output was compared to previously calculated surface values of folded globular proteins.[7] The results show very good agreement between the two sets of data (Figure 2). Then, surface values in the unfolded state were calculated by ProtSA for all the proteins considered in this study. A double-log plot for the charge-to-surface relation was then obtained (Figure 3), similar to the previously published analysis on folded proteins.[7] The results show considerable similarity in the trend observed for folded and unfolded proteins. As above, IDPs under nondenaturing conditions are used as a control for solvent effects. Different from the charge-to-mass relation,

the charge-to-surface relation seems to hold for either folded or unfolded proteins. There is a small discrepancy in the slope values derived from the two sets of proteins, but the agreement is quite good, particularly for low to medium molecular weights. It cannot be ruled out that the employed algorithm underestimates slightly but systematically the surface area for large proteins in the denatured state. Nevertheless, these results strongly suggest that the extent of protein ionization by electrospray correlates with the solvent-accessible surface area, independent of the conformational state. An implication of these results is that ESI-MS data can be used to estimate the solvent-accessible surface area and, hence, to formulate hypotheses on the fold state of unknown proteins. For instance, IDPs will give estimated surface areas much above the typical values for folded globular proteins of the same mass.

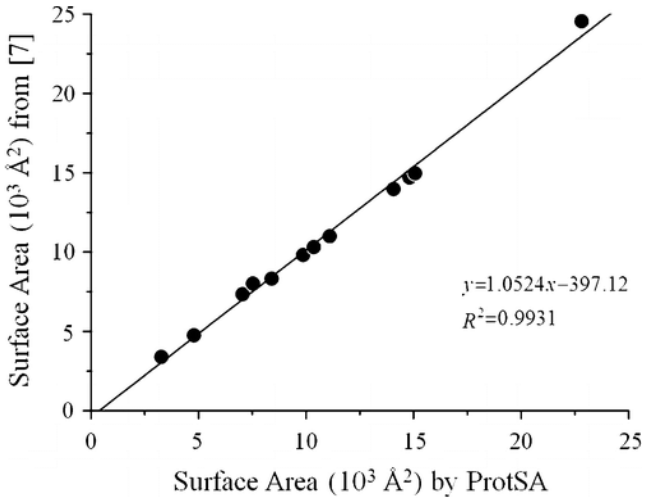


Figure 2. Validation of ProtSA calculations. The previously reported[7] solvent-accessible surface area of folded proteins is plotted against ProtSA output. The PDB codes were given as input for the calculations. The equation and R-squared value for the linear regression are indicated on the figure.

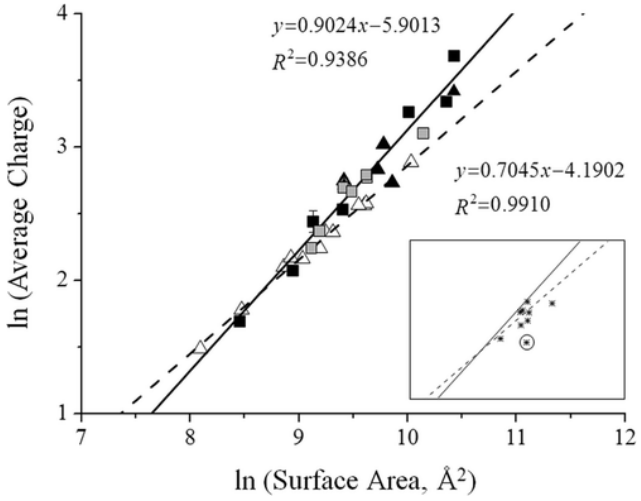


Figure 3. Charge-to-surface correlation for globular proteins and IDPs analyzed under nondenaturing and denaturing conditions. The logarithm of the experimental average charge in the positive-ion mode is plotted against the logarithm of the solvent-accessible surface area calculated by ProtSA. Black squares, IDPs under denaturing conditions. Black triangles, globular proteins under denaturing conditions. Gray squares, IDPs under nondenaturing conditions (in the case of multimodal distributions, only the high-charge component was considered). The latter points are overlaid to the plot as a control for solvent effects and were not used to calculate the linear regression. Open triangles, globular proteins under nondenaturing conditions. The ProtSA values for unfolded proteins are the average of three independent runs. The standard deviation is within 0.02%. Therefore, horizontal error bars would be smaller than the symbol size. Vertical error bars are given only for the data obtained in our laboratory and represent the standard deviation from three independent experiments. The fitting is shown as a continuous line for proteins under denaturing conditions and as a dashed line for globular proteins under nondenaturing conditions. The equations and R-squared values for the linear regression are indicated on the figure. The inset shows data points from measurements in the negative-ion mode (asterisks), relatively to the linear fitting for the positive-ion mode. The lysozyme point is circled.

The inset in Figure 3 shows some points from the negative-ion mode. In this case, proteins are deprotonated during the electrospray process and only ions with a negative net charge can be detected. It is generally more difficult to ionize and detect proteins in negative ion mode, although the reason for that is still unclear. Therefore, fewer data are available in the literature. As can be seen in the figure, these points are more scattered than those from the positive-ion mode. Furthermore, there is a general tendency to lie at lower values of average charge. However, these points show a similar trend to that

observed in positive-ion mode. It is interesting to note that a clear outlier is represented by lysozyme, a very basic protein (pI 9.36). This result suggests that protein ionization in negative-ion mode might also become limited by unfavorable amino acid composition. The disulfide bonds that characterize the native lysozyme structure should not be the reason, since the protein was analyzed under reducing conditions (Table 1).

The charge-to-mass relation of folded, globular proteins is usually interpreted by the Rayleigh-limit hypothesis.[4-6] According to this hypothesis, the charge acquired by a protein during electrospray reflects the charge of the precursor ESI droplet, which in turn, is considered to be close to the Rayleigh limit. Supporting this interpretation, the experimental points can be fit well by the Rayleigh equation for water droplets of the same size as the protein structures.[4-6] On the other hand, three problems weaken this interpretation. (i) The agreement with the Rayleigh equation breaks down when solvents with different surface tension than water are considered.[27,34] (ii) The intrinsic net charge of the protein in solution is implicitly required to always be zero, since its final charge is considered to be equal to the number of protons acquired from the ESI droplet. This assumption cannot hold in general since proteins are often analyzed at pH values distant from their pI. (iii) A different mechanism has been suggested to govern ionization of unfolded proteins, i.e., the apparent gas-phase basicity (GBapp) of the protein relative to the GB of the solvent.[35] GB is the enthalpy change associated with proton-transfer reactions in the gas phase and allows interpreting the behavior of many different molecules under electrospray conditions, including polypeptides.[36] The present results could be consistent with GB being the general limiting factor for protein ionization, independent of the conformational state. Indeed, charge density on the protein surface determines the extent of electrostatic shielding of protonable residues (mainly neutral basic residues and negatively charged carboxylates), modulating their GBapp within the polypeptide chain.[17, 36, 37]

The general charge-surface correlation pointed out helps rationalize conformational effects in protein ESI-MS. Indeed, protein unfolding causes a dramatic increase in the solvent-accessible surface area, implying higher protein ionization according to the above relation. A complex network of intramolecular interactions likely characterizes protein ions produced by electrospray, including electrostatic repulsions due to protein ionization but possibly also attractive long-range interactions, ion pairs, hydrogen bonds, and cation−π interactions involving protonable residues. All these interactions take place mostly on the protein surface and could result into the observed correlation of protein charge with a solvent-accessible surface area.

Thus, although protonable residues are generally exposed on the protein surface in either folded or unfolded proteins, their interactions change with protein conformation and, ultimately, with the protein surface. This concept extends the so-called conformation-dependent neutralization (CDN) hypothesis,[37] which suggested that zwitterionic states might survive more easily in folded than unfolded proteins, contributing to conformational effects in ESI-MS. This could be one mechanism mediating the charge-to-surface correlation reported here. It should be noted that the CDN hypothesis has been erroneously interpreted to imply that the observed charge states should correspond in general to the difference between basic and acidic residues,[5] which was only supposed to be the limit case for basic proteins in the positive ion mode and acidic proteins in the negative ion mode. Computational approaches could help rationalizing in terms of GBapp the ionization of folded and unfolded proteins and the differences between the positive- and negative-ion modes.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rita.grandori@unimib.it.
Phone: +390264483363.

ACKNOWLEDGMENT

The authors wish to thank Sonia Longhi for providing the plasmids for NTAIL and PNT expression. This work was supported by the grants FAR ("Fondo Ateneo per la Ricerca") of Milano-Bicocca University to R.G. and S.B. and by a doctoral fellowship "Fondazione Fratelli Confalonieri" to L.T.

REFERENCES

(1) Sharon, M.; Robinson, C. V. Annu. Rev. Biochem. 2007, 76, 167–193.
(2) Kaddis, C. S.; Loo, J. A. Anal. Chem. 2007, 79, 1778–1784.
(3) Kebarle, P.; Verkerk, U. H. Mass Spectrom. Rev. 2009, 28, 898–917.
(4) De la Mora, J. F. Anal. Chim. Acta 2000, 406, 93–104.
(5) Nesatyy, V. J.; Suter, M. J. J. Mass Spectrom. 2004, 39, 93–97.
(6) Heck, A. J.; Van Den Heuvel, R. H. Mass Spectrom. Rev. 2004, 23, 368–389.
(7) Kaltashov, I. A.; Mohimen, A. Anal. Chem. 2005, 77, 5370–5379.
(8) Frimpong, A. K.; Abzalimov, R. R.; Uversky, V. N.; Kaltashov, I. A. Proteins 2010, 78, 714–722.

(9) Hautreux, M.; Hue, N.; Du Fou de Kerdaniel, A.; Zahir, A.; Malec, V.; Laprevote, O. Int. J. Mass Spectrom. 2004, 231, 131–137.
(10) Chowdhury, S. K.; Katta, V.; Chait, B. T. J. Am. Chem. Soc. 1990, 112, 9012–9013.
(11) Karlin, D.; Longhi, S.; Receveur, V.; Canard, B. Virology 2002,296, 251–262.
(12) Bourhis, J. M.; Receveur-Brechot, V.; Oglesbee, M.; Zhang, X.;Buccellato, M.; Darbon, H.; Canard, B.; Finet, S.; Longhi, S. Protein Sci. 2005, 14, 1975–1992.
(13) Brocca, S.; Testa, L.; Samalikova, M.; Grandori, R.; Lotti, M. Mol. Biotechnol. 2011, 47, 34–42.
(14) Brocca, S.; Testa, L.; Sobott, F.; Samalikova, M.; Natalello, A.; Papaleo, E.; Lotti, M.; De Gioia, L.; Doglia, S. M.; Alberghina, L.; Grandori, R. Biophys. J. 2011, 100, 2243–2252.
(15) Santambrogio, C.; Ricagno, S.; Colombo, M.; Barbiroli, A.; Bonomi, F.; Bellotti, V.; Bolognesi, M.; Grandori, R. Protein Sci. 2010, 19, 1386–1394.
(16) Estrada, J.; Bernado, P.; Blackledge, M.; Sancho, J. BMC Bioinf. 2009, 10, 104.
(17) Konermann, L. J. Phys. Chem. B 2007, 111, 6534–6543.
(18) Canon, F.; Pate, F.; Meudec, E.; Marlin, T.; Cheynier, V.; Giuliani, A.; Sarni-Manchado, P. Anal. Bioanal. Chem. 2009, 395, 2535–2545.
(19) Murphy, P. W.; Rowland, E. E.; Byers, D. M. J. Am. Soc. Mass Spectrom. 2007, 18, 1525–1532.
(20) Uchiki, T.; Hettich, R.; Gupta, V.; Dealwis, C. Anal. Biochem. 2002, 301, 35–48.
(21) Yi, S.; Boys, B. L.; Brickenden, A.; Konermann, L.; Choy, W. Y. Biochemistry 2007, 46, 13120–13130.
(22) Testa, L.; Brocca, S.; Samalikova, M.; Santambrogio, C.; Alberghina, L.; Grandori, R. Biotechnol. J. 2011, 6, 96–100.
(23) Iakoucheva, L. M.; Kimzey, A. L.; Masselon, C. D.; Bruce, J. E.; Garner, E. C.; Brown, C. J.; Dunker, A. K.; Smith, R. D.; Ackerman, E. J. Protein Sci. 2001, 10, 560–571.
(24) Brocca, S.; Samalikova, M.; Uversky, V. N.; Lotti, M.; Vanoni, M.; Alberghina, L.; Grandori, R. Proteins 2009, 76, 731–746.
(25) Samalikova, M.; Matecko, I.; Muller, N.; Grandori, R. Anal. Bioanal. Chem. 2004, 378, 1112–1123.
(26) Konermann, L.; Douglas, D. J. J. Am. Soc. Mass Spectrom. 1998, 9, 1248–1254.
(27) Samalikova, M.; Grandori, R. J. Mass Spectrom. 2005, 40, 503– 510.
(28) Invernizzi, G.; Casiraghi, L.; Grandori, R.; Lotti, M. J. Biotechnol. 2009, 141, 42–46.
(29) Watt, S. J., Ph.D. Thesis, Department of Chemistry, University of Wollongong, Australia, 2005,http://ro.auo.edu.au/theses/582.
(30) Natalello, A.; Benetti, F.; Doglia, S. M.; Legname, G.; Grandori, R. Proteins 2010, 79, 611–621.
(31) Pan, J.; Konermann, L. Biochemistry 2010, 49, 3477–3486.
(32) Sogbein, O. O.; Simmons, D. A.; Konermann, L. J. Am. Soc. Mass Spectrom. 2000, 11, 312–319.
(33) Veenstra, T. D.; Johnson, K. L.; Tomlinson, A. J.; Kumar, R.; Naylor, S. Rapid Commun. Mass Spectrom. 1998, 12, 613–619.

(34) Samalikova, M.; Grandori, R. J. Am. Chem. Soc. 2003, 125, 13352–13353.
(35) Schnier, P. D.; Gross, D. S.; Williams, E. R. J. Am. Soc. Mass Spectrom. 1995, 6, 1086–1097.
(36) Marchese, R.; Grandori, R.; Carloni, P.; Raugei, S. PLoS Comput. Biol. 2010, 6, e1000775.
(37) Grandori, R. J. Mass Spectrom. 2003, 38, 11–15.

Defining Structural Domains of an Intrinsically Disordered Protein: Sic1, the Cyclin-Dependent Kinase Inhibitor of *Saccharomyces cerevisiae*

Molecular Biotechnology. 2011 Jan;47(1):34-42.

Stefania Brocca*, Lorenzo Testa, Maria Šamalikova, Rita Grandori and Marina Lotti

Department of Biotechnology and Biosciences, Università di Milano-Bicocca, P.zza della Scienza 2, 20126 Milan, Italy

* Corresponding author
E-mail: stefania.brocca@unimib.it

ABSTRACT

The cyclin-dependent kinase inhibitor Sic1 is an intrinsically disordered protein (IDP) involved in cell–cycle regulation in the yeast Saccharomyces cerevisiae. Notwithstanding many studies on its biological function, structural characterization has been attempted only recently, fostering the development of production and purification protocols suitable to yield large amounts of this weakly expressed protein. In this study, we describe the identification of protein domains by the heterologous expression, purification, and characterization of Sic1-derived fragment. Four C-terminal fragments ($Sic1^{C\text{-}ter}$) were produced based on functional studies and limited-proteolysis results. The N-terminal fragment ($Sic1^{1–186}$) was complementary to the most stable C-terminal fragments ($Sic1^{\Delta 186}$). Both $Sic1^{1–186}$ and $Sic1^{C\text{-}ter}$ fragments were, in general, less susceptible to spontaneous proteolysis than the full-length protein. The boundaries of the C-terminal fragments turned out to be crucial for integrity of the recombinant proteins and required two rounds of design and production. Sic1 fragments were purified by a simple procedure, based on their resistance to heat treatment, at the amount and purity required for structural characterization. Circular dichroism (CD) measurements and nuclear magnetic resonance (NMR) spectra of N- and C-terminal fragments confirm their disordered nature but reveal minor structural differences that may reflect their distinct functional roles.

Keywords

Escherichia coli - Disordered domains - Heterologous expression - Mass spectrometry - Circular dichroism

Introduction

Intrinsically disordered proteins (IDPs) are defined by the lack of compact, three-dimensional structure under physiological conditions, in the absence of specific interactors. IDPs, in fact, acquire an ordered structure only upon binding to interaction partners or by functional regulation via post-translational modifications [1–5]. The prominent feature of IDPs is, therefore, their structural flexibility and pronounced conformational dynamics, although they can transiently visit structured or compact states [6]. The literature on IDPs has flourished during the last decade, witnessing a great deal of interest on structural disorder and pointing out its physiological importance [7–9]. IDPs are frequently involved in key biological processes such as cell–cycle control, transcription and translation regulation, membrane fusion and transport, and signal transduction [10, 11]. A high percentage of cell-signaling and cancer-associated proteins are predicted to have long disordered regions [12]. These examples indicate a general importance of intrinsic disorder for signaling and regulation [13].

The protein Sic1 plays a key role in regulating the transition from the G1 to S phase during the mitotic cell cycle in the yeast, Saccharomyces cerevisiae [14]. This function is accomplished through the formation of ternary complexes with the Cdk1 kinase and its cognate cyclin (Clb5 or Clb6), coordinating the major events of cell-cycle progression in S. cerevisiae [15]. Kinase activity is inhibited by the formation of the Sic1–Cdk1–Clb5/6 complex, and restored after ubiquitin-dependent Sic1 degradation, which is, in turn, triggered by Sic1 phosphorylation at multiple sites in its N-terminal region [16, 17]. The kinase-inhibitory domain (KID) has been identified in the last 70 amino acids, as the minimal fragment required for in vivo inhibition of the Cdk1–Clb5/6 activity (aa 215–284) [18].

Although Sic1 has been the object of accurate experimental and theoretical study concerning the effects of phosphorylation on regulatory pathways [19–21], its structural characterization has been undertaken only very recently. Bioinformatics analysis and limited-proteolysis experiments have shown that the N-terminus is the most disordered part of the protein, while the C-terminus is characterized by higher structural compactness [22]. Proteolysis-hypersensitive sites cluster at the N-terminus and in the central part of the

protein [22]. The molecule is characterized by remarkable compactness, making it more similar to a molten globule than to a fully extended protein [22]. A fragment corresponding to the first N-terminal 90 residues has been studied by nuclear magnetic resonance (NMR) [23], revealing significant amount of transient secondary and tertiary structure. Altogether, these studies suggest that, in spite of the disordered nature of this protein, it is possible to recognize distinct "structural domains". Such a modular organization of the Sic1 polypeptide could be consistent with the functional and evolutionary features of its KID domain [18, 22, 23]. We, therefore, undertook expression and purification of $Sic1^{N\text{-}ter}$ and $Sic1^{C\text{-}ter}$ fragments, with the scope to define protein domains sufficiently stable to be characterized by circular dichroism (CD) and NMR studies.

Materials and Methods

Strains and Growth Media

Escherichia coli strain DH5α (Invitrogen, Life Technologies Corporation, Carlsbad, California) was used as the host for DNA amplification, whereas strain BL21-Rosetta 2(DE3) (Novagen, EMD Chemicals Inc, Darmstadt, DE) was used for heterologous expression.
Escherichia coli cells were grown in low-salt Luria–Bertani (ls-LB) medium (10 g peptone, 5 g yeast extract, and 5 g NaCl in 1 l water). E. coli transformants were selected on ls-LB agar plates supplemented with the appropriate antibiotics [34 mg/l chloramphenicol and 50 mg/l ampicillin for BL21-Rosetta 2(DE3), and 50 mg/l ampicillin for DH5α].

Cloning

Standard recombinant-DNA techniques were applied according to Sambrook et al. [24]. Restriction enzymes and DNA ligase were from New England Biolabs (NEB, Ipswich, MA, USA), while the primers were purchased from Primm (Milano, Italy) or Sigma-Aldrich (St. Louis, MO, USA). The SIC1 encoding sequence from pET21[$SIC1^{HIS6}$] [22] was subjected to mutagenic PCR to obtain deletions of its 5′- or 3′-moieties. Mutagenesis was carried out by *back-to-back* PCR [25] with non-overlapping phosphorylated oligonucleotides (to produce $SIC1^{C\text{-}ter}$) or with overlapping non-phosphorylated oligonucleotides (to produce $SIC1^{1–186}$). Both kinds of primers were designed to amplify the whole plasmid containing the target gene, excluding the DNA fragment to be deleted. Products obtained by *back-*

to-back PCR with overlapping oligonucleotides contained terminal restriction sites generating, upon digestion, phosphorylated, compatible ends suitable to circularize the amplimers by self-ligation. In this case, overlapping non-phosphorylated primers were designed to contain, at their 5′-terminus, a recognition site for the *MscI* restriction enzyme (underlined in Table 1), preceded by a four-base extension (boxed in Table 1) to improve cleavage efficiency. After PCR, the amplimers were digested by *MscI* and then circularized by self-ligation reconstituting a complete restriction site for *MscI*. The insertion of the MscI site at the end of the Sic1 open reading frame introduced additional codons for one glycine and one histidine residue. Therefore, a seven-histidine tag was obtained.

When non-overlapping, phosphorylated oligonucleotides were used, the blunt-end amplimers were directly ligated resulting in circularization of the plasmidic DNA. In particular, each of the non-overlapping phosphorylated primers was designed to contain at its 5′-terminus a triplet corresponding to the *NdeI* half site (shaded in Table 1). After PCR, amplimer circularization through self-ligation reconstituted a complete, functional restriction site for *NdeI*, used for the screening of mutated plasmids. Oligonucleotides phosphorylation was carried out before PCR, by incubation with polynucleotide kinase A (NEB, Ipswich, MA, USA) at 37°C for 1 h. Deletions were obtained by PCR with 10 ng of pET21[SIC1^{HIS6}] as a template, with 0.5 μM of each primer. Sequences of the oligonucleotides employed in this study are reported in Table 1.

PCR amplification was carried out in a volume of 50 μl using the high-fidelity *PfuII* Ultra DNA polymerase (Stratagene, La Jolla, CA) according to the manufacturer's instructions and applying the following temperature program: 2-min denaturation at 95°C, addition of 1 μl DNA polymerase, 30 cycles of 30 s at 95°C/30 s at 50–60°C/4 min at 72°C, and a final extension step of 10 min at 72°C. The template DNA was digested by *DpnI* after amplification. The amplified DNA, as obtained (when amplified with non-overlapping oligonucleotides) or digested by *MscI* (when amplified with overlapping oligonucleotides), was purified by ethanol precipitation before self-ligation by the T4 DNA ligase. Ligations were carried out with 50 ng DNA in 20 μl reaction mixture at 14°C.

The recombinant plasmids were introduced into DH5α or BL21-Rosetta 2(DE3) *E. coli* cells, according to a procedure adapted from [26], as described by [27]. The final constructs were checked by DNA sequencing (Primm, Milano, Italy).

Oligonucleotide sequences	Construct	Sic1 fragments
Reverse C-ter 5'-ATG TAT ATC TCC TTC TTA AAG TTA AAC AAA ATT ATT TCT AG -3'		
Forward Δ171 5'- ATG GGT ACG CCC AGC GAC AAG GTG -3'	pET21a [SIC1$^{\Delta 171}$]	Sic1$^{172-284}$
Forward Δ175 5'- ATG GAC AAG GTG ATA ACA TTT GAA TGG GC -3'	pET21a [SIC1$^{\Delta 175}$]	Sic1$^{176-284}$
Forward Δ186 5'- ATG TGG AAC AAC AAC TCT CCG AAA AAT CAG C -3'	pET21a [SIC1$^{\Delta 186}$]	Sic1$^{187-284}$
Forward Δ214 5'- ATG GGT AAA AAT CCC TTT GCA TCA GAT G -3'	pET21a [SIC1$^{\Delta 214}$]	Sic1$^{215-284}$
Forward N-ter 5'- ACGT TGG CCA TTT TTT GCC AAT TCA AAT G -3'		
Reverse 1-186 5'- ACGT TGG CCA ACC CGC ACT GGA TTG ATG ATG -3'	pET21a [SIC1$^{1-186}$]	Sic1$^{1-186}$

Table1. Primers and vectors used for production of Sic1 fragments.
The deletions of the 5' region of SIC1 for production of Sic1$^{C\text{-}ter}$ were obtained by PCR with different forward primers and the same reverse *C-ter* primer. The deletion of the 3' region of SIC1 was obtained with the forward *N-ter* primer and the reverse 1–186 primer. *Shaded sequences* indicate the *NdeI* half site in primers used for deletion of the SIC1 5' region. *Underlined sequences* indicate the *MscI* site in the primers used for deletion of the SIC1 3' region. *Boxed sequences* indicate the extensions inserted on overlapping oligonucleotides to improve their restriction efficiency.

Protein Expression and Purification

Escherichia coli pre-cultures were inoculated from a single-cell colony from a fresh selection plate. Preparative cultures were obtained in shaking flasks by 1:20 dilution of pre-cultures in fresh ls-LB medium volumes scaled from 20 to 200 ml and grown at 37°C and 220 rpm to an optical density at 600 nm (OD600) between 0.5 and 0.7. Cultures were then transferred to 30°C (to reduce the risk of proteolytic degradation) and induced with 200 µM isopropyl-β-d-thiogalactopyranoside (IPTG).
Induced cells were harvested by centrifugation at 9,600 g at room temperature for 15 min and either immediately used for protein extraction or frozen at −20°C. The cell pellet was resuspended in extraction buffer (50 mM sodium phosphate, 300 mM NaCl, 10 mM imidazole, pH 8.0) containing a protease-inhibitor cocktail for use in purification of histidine-tagged proteins (Sigma-Aldrich, St. Louis, MO, USA), and cells were lysated by incubating the cell suspension at 99°C for 20 min. The crude extracts were then cooled for 10 min on ice and centrifuged for 10 min at 10,600 g to collect cellular debris and the insoluble protein fraction. The

supernatant was loaded on an immobilized-metal affinity chromatography (IMAC) on Ni^{2+}/NTA beads, as previously described [28]. The column was eluted by 50 mM sodium phosphate, 300 mM NaCl, 250 mM imidazole, pH 8.0 [22]. For comparison, crude extracts by sonication instead of heat treatment were prepared, too. In this case, cells were lysated by eight 15-s pulses on a Fisher sonicator model 300 (Artec Systems Corp, 170 Finn Court, Farmingdale, NY: U.S.A.), with relative output 50%.
The purified proteins were exchanged to the final buffer (50 mM sodium phosphate, pH 6.5 or 50 mM ammonium acetate, pH 6.5) by two consecutive gel filtrations on PD-10 columns (General Electric Healthcare, Little Chalfont, UK). Protein concentration was determined by the Bradford protein assay (Bio-Rad Laboratories, Milano, Italy), using bovine serum albumin as a standard.

SDS-PAGE

SDS-PAGE analyses were carried out on 16% acrylamide Laemmli gels [29]. After electrophoresis, gels were stained by GelCode Blue (Pierce, Illinois, IL, USA). Broad-range, prestained molecular-weight markers (NEB, Ipswich, MA, USA) were used as standards.

Spectroscopy

The CD spectra were recorded on a spectropolarimeter J-815 (JASCO corporation, Japan), in a 1-mm pathlength cuvette, at 25°C. Samples were in 50 mM sodium phosphate buffer, pH 6.5. The spectra were acquired with data pitch 0.2 nm, averaged over three acquisitions, and smoothed by the Savitzky-Golay algorithm [30].
The NMR spectra were recorded on a Bruker Avance III 600 spectrometer equipped with a pulsed field gradient triple resonance TCI cryoprobe. Standard sequence schemes employing pulsed field gradients were used to achieve suppression of the solvent signal and spectral artifacts. Direct and indirect dimensions were normally apodized by the use of 90°-shifted squared sine-bell functions or Lorentzian-to-Gaussian functions, followed by zero filling and Fourier transform. TOCSY spectra were recorded employing isotropic mixing times in the range 80–120 ms, while NOESY spectra were obtained for different mixing times ranging from 120 to 300 ms. The NMR data were processed and analyzed with Topspin 2.1 (Bruker). Typical samples contained 0.3 mM protein dissolved in 150 mM NaCl, 90% H2O, and 10% D_2O. Spectra were acquired at 298 and 283 K.

Results and Discussion

Expression and Purification of Sic1 C-Terminal Fragments

Structural characterization of full-length Sic1 by limited proteolysis and mass spectrometry [22] localized its most compact domain in the C-terminal region. Since the fragments showing the highest relative resistance to proteases started at position 172 and 176 [22], fragments 172–284 ($Sic1^{\Delta 171}$) and 176–284 ($Sic1^{\Delta 175}$) were subcloned. In addition, fragments 215–284 ($Sic1^{\Delta 214}$), corresponding exactly to the previously identified functional KID domain [31], were also produced (Fig. 1). The recombinant, histidine-tagged proteins were expressed from pET vectors in BL21-RosettaTM 2(DE3) *E. coli* cells.

Soluble proteins extracted from induced cells taken at 1, 2, and 4 h after induction were analyzed by SDS-PAGE, showing the already accumulated recombinant proteins 1 h after induction. SDS-PAGE analyses showed that C-terminal fragments are produced at higher amounts than the full-length protein (data not shown).

Preparative cultures were induced for 2 h. Proteins were extracted from fresh cell pellets. For protein purification, we took advantage of the fact that aggressive denaturing agents, like heat and acids, have often only minor and reversible effects on intrinsically disordered proteins, while inducing aggregation and precipitation of normally folded proteins [22, 32]. This property has been successfully exploited for IDP enrichment in proteomic studies [33–35]. Heat treatment can also be exploited for cell lysis. This procedure offers the additional advantage, relevant for IDP production, of reducing the extraction time and minimizing the exposure to proteolytic enzymes. Therefore, induced cells were disrupted by incubation at 99°C for 20 min. After removal of the insoluble material by centrifugation, the sample was further purified by IMAC. The results of this purification procedure are shown in Fig. 1. Gel densitometry indicates a purity degree above 95% of the final samples. Additional bands, especially in $Sic1^{\Delta 171}$ and $Sic1^{\Delta 175}$ samples, are derived from proteolytic degradation of target proteins, as demonstrated by western blot analyses (data not shown).

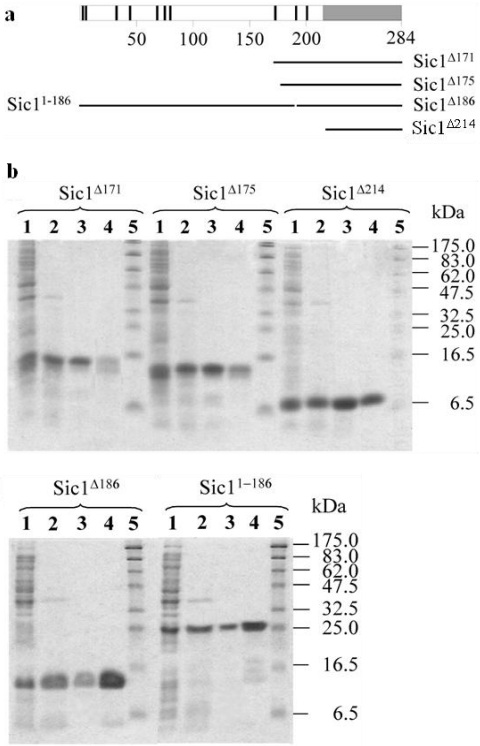


Figure 1. Production of Sic1 fragments. **a** Schematic representation of the Sic1 molecule and the N- and C-terminal fragments produced in the course of this study. Known phosphorylation sites are indicated by black bars, and the functional KID domain is shaded. Each C-terminal fragment ($Sic1^{C\text{-}ter}$) was designed as ''Δ number'', with the number indicating the last amino acid deleted by mutagenesis. The N-terminal fragment spanning the amino acids from 1 to 186 ($Sic1^{1–186}$) was designated as $Sic1^{1–186}$. **b** Purification of Sic1 fragments analyzed by Coomassie-stained SDS-PAGE gels. *Lane 1* soluble crude extract, *lane 2* heat-treated, soluble crude extract (sample loaded on IMAC), *lane 3* IMAC-purified protein, *lane 4* fraction eluted from gel filtration, *lane 5* molecular-weight markers.

The purified proteins were finally subjected to buffer exchange by gel filtration according to the requirements of subsequent analyses. Low ionic strength is generally required, although with different levels of salt tolerance, by biophysical techniques aimed at structural characterization, such as CD, electrospray-ionization mass spectroscopy (ESI–MS), and Fourier transform infrared (FTIR) spectroscopy. However, Sic1 and its fragments are poorly soluble in pure water (data not shown). The pure proteins could be obtained at concentrations around 100 μM in 50 mM phosphate or acetate buffer at pH 6.5.

Figure 2 shows the deconvoluted ESI–MS spectra of 10 μM protein samples in 50% acetonitrile, 1% formic acid. Mass deconvolution of $Sic1^{\Delta171}$ (Fig. 2a) reveals a main peak of 14147 ± 2 Da, corresponding to the calculated mass of the Sic1 172–284 fragment fused with the histidine tag and depleted of the initial methionine (14146.5 Da). On the contrary, $Sic1^{\Delta175}$ (Fig. 2b) has a main peak at 13936 ± 2 Da, corresponding to the calculated mass of the fragment with tag and initial methionine (13935.4 Da). According to Frottin and co-workers [36], the N-terminal methionine of proteins expressed in *E. coli* is cleaved off only if followed by a residue with a small side chain, such as glycine, alanine, serine, and proline. The observed pattern is in agreement with this rule, since the residue following the initial methionine is glycine in $Sic1^{\Delta171}$ and aspartic acid in $Sic1^{\Delta175}$. The $Sic1^{\Delta214}$ spectrum (Fig. 2) reveals two peaks, a major one at 9294 ± 2 Da, corresponding to the calculated mass of the fragment with histidine tag, without the initial methionine (9293 Da), and a minor one at 9425 ± 2 Da, corresponding to the calculated mass of the fragment with histidine tag and N-terminal methionine (9425 Da). The second aminoacid in $Sic1^{\Delta214}$ is glycine, which explains the prevalence of the cleaved product. Residual protein retaining the initial methionine even in a sequence context that favors its removal is likely due to protein overexpression.

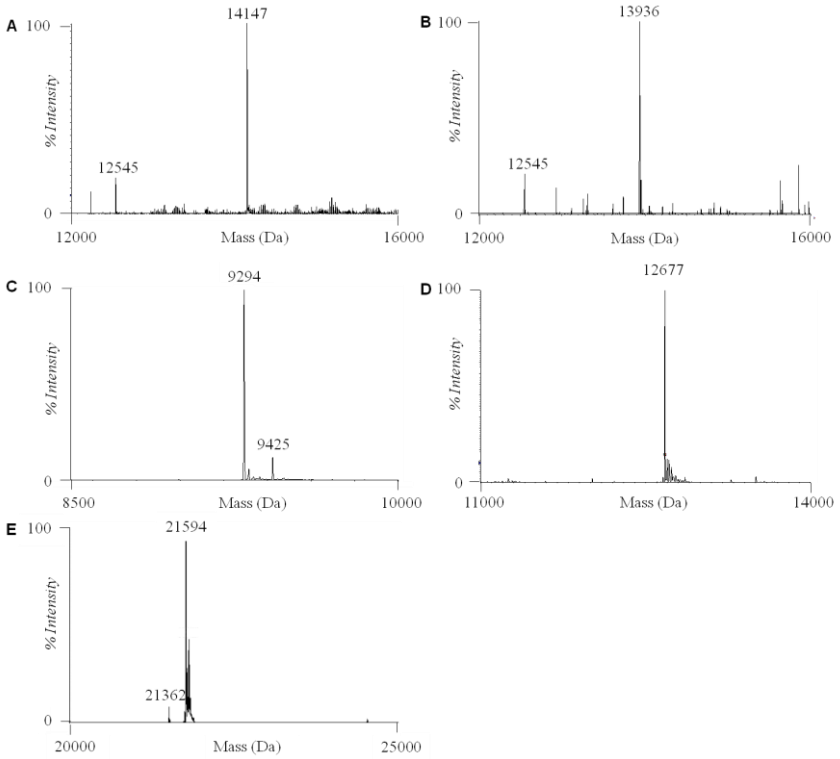


Figure 2. Nano-ESI–MS deconvolution spectra of 10 μM Sic1$^{\Delta 171}$ (**a**), Sic1$^{\Delta 175}$ (**b**), Sic1$^{\Delta 214}$ (**c**), Sic1$^{\Delta 186}$ (**d**), Sic1$^{1-186}$ (**e**). Samples were sprayed from solutions containing 1% formic acid, 50% acetonitrile

In the spectra of Sic1$^{\Delta 171}$ and Sic1$^{\Delta 175}$ (Fig. 2a, b), a secondary peak at 12545 ± 2 Da was observed, suggesting the loss of the first 15 and 11 residues, respectively, generating the Sic1 fragment 187–284 fused with the histidine tag (calculated mass 12544.7 Da). These results imply that the Sic1$^{\Delta 171}$ and Sic1$^{\Delta 175}$ polypeptides are unstable at their N-terminus and consistently indicate propensity to generate the shorter Sic1$^{C\text{-ter}}$ fragment starting from residue 187. Mass heterogeneity in the final preparations of Sic1$^{\Delta 171}$ and Sic1$^{\Delta 175}$ is clearly visible also from SDS-PAGE (Fig. 1). Therefore, the Sic1 fragment 187–284 (Sic1$^{\Delta 186}$) was subcloned, too, with the aim to obtain a stable Sic1$^{C\text{-ter}}$ fragment that most closely matches the late fragments identified by limited proteolysis. Mass deconvolution of Sic1$^{\Delta 186}$ (Fig. 2d) shows a unique peak at 12677 ± 2 Da, corresponding to the calculated mass (12675.9 Da) of the protein with histidine tag and initial methionine (the second residue is tryptophan). No evidence of degradation of such fragment emerges by mass spectrometry or SDS-PAGE. Therefore, Sic1$^{\Delta 186}$ and

$Sic1^{\Delta 214}$ were selected as the relevant $Sic1^{C\text{-}ter}$ fragments, representing the C-terminal domain according, respectively, to a more structural or functional definition of protein domain, as derived from limited-proteolysis and in vivo complementation studies.

Expression and Purification of a Sic1 N-Terminal Fragment

The N-terminal fragment has been designed as the complementary fragment to $Sic1^{\Delta 186}$ and contains, therefore, residues 1–186 ($Sic1^{1–186}$) (Fig. 1). This fragment was expressed at lower levels (~10 mg/l) than $Sic1^{C\text{-}ter}$ fragments (~30 mg/l) (Fig. 1). This difference can be due to higher stability of C-terminal fragments against proteolytic degradation. The protein was extracted and purified after 2-hour induction, according to the protocol already described (Fig. 1b).
Deconvoluted ESI–MS spectrum of $Sic1^{1–186}$ shows a main peak at 21594 ± 2 Da (Fig. 2e), corresponding to the mass of the fragment with histidine tag and initial methionine (21592.97 Da), as expected being the second amino acid a threonine. The small peak at 21362 ± 2 Da in Fig. 2e indicates loss of the first two residues, Met1 and Thr2, yielding the fragment 3–186 (calculated mass 21360.7 Da).

Structural Characterization by CD and NMR Spectroscopy

The CD spectra of IDPs are dominated by the characteristic features of random-coil polypeptides, with a strong negative band around 200 nm and still negative ellipticity values at 190 nm [37]. The complementary fragments $Sic1^{\Delta 186}$ and $Sic1^{1–186}$ were analyzed by CD spectroscopy in the far UV, in order to probe their secondary-structure properties. The spectra of the two fragments are compared in Fig. 3. In both cases, the mean-residue ellipticity is characterized by a negative minimum around 200 nm and a negative shoulder at 222 nm. These results indicate that the isolated fragments are mainly disordered with a very small content in alpha-helix, similarly to the full-length protein [22]. However, the C-terminal fragment seems to be slightly more structured than the N-terminal one, as indicated by lower ellipticity values at 190 and 200 nm and by a minor shift of the negative minimum toward higher wavelengths (Fig. 3). This result is consistent with the evidence of stronger order propensity in the Sic1 C-terminal region [22].

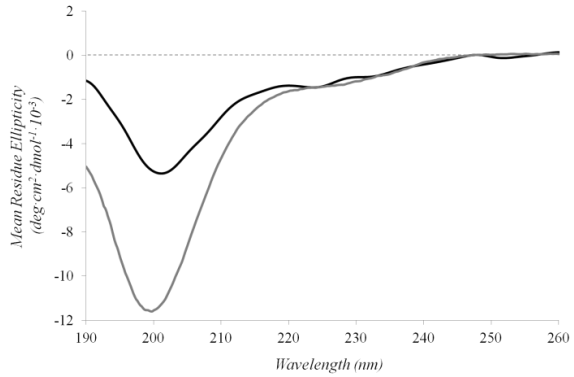


Figure 3. CD spectra in the far UV of 10 μM Sic1$^{1–186}$ (gray) and 10 μM Sic1$^{\Delta186}$ (black) in 50 mM sodium phosphate pH 6.5

The NMR spectroscopy is one of the most appropriate techniques to offer detailed insight into disordered states [38]. Therefore, NMR investigation of the Sic1$^{\Delta186}$ fragment was also attempted. Unfortunately, the solubility of the protein in the CD buffer was too low to reach concentrations suitable for NMR. In order to overcome this problem, solubility tests in phosphate buffer at increasing NaCl concentrations were performed (data not shown). The protein could be concentrated up to 300 μM in 150 mM NaCl. Such samples were used to record 1H NMR spectra. The 1D 1H spectrum is shown in Fig. 4a, the aromatic region of the TOCSY spectrum in Fig. 4b, while Fig. 4c shows a superposition of the NH–NH$_{i+1}$ region of the NOESY spectra recorded at 283 and 298 K. The poor peak dispersions observed in all the spectra confirm the intrinsically disordered nature of this protein fragment. No residue-specific chemical-shift assignment could be obtained from the data. Aromatic resonances are indicated in Fig. 4b, according to the type of residue. They were identified on the basis of the coupling pattern, the typical random-coil shift, and the presence of NOE cross peaks between aromatic and β protons at their typical positions. NOESY spectra recorded at 283 K, to investigate whether more NH–NH$_{i+1}$ correlations could be extracted at a lower temperature, did not exhibit a more extended pattern of correlations, which could have highlighted a population with residual structure, as reported, for example, for the BNP peptide [39].

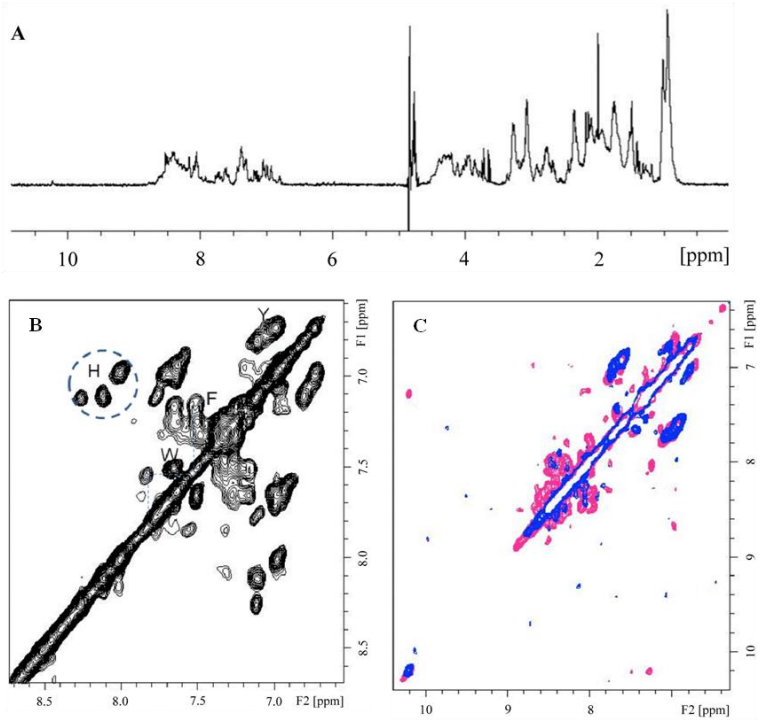


Figure 4. 600 MHz 1H NMR spectra of 300 μM Sic1$^{\Delta 186}$ in 150 mM NaCl. **a** 1D spectrum, **b** 2D-NMR TOCSY spectrum of the aromatic region, reporting the assignment of the aromatic residue types, **c** superposition of the 2D-NMR NOESY spectra for the NH–NH_{i+1} region obtained at 25°C (black) and 10°C (gray)

Conclusions

The procedures reported here enable production of Sic1 fragments at amounts and yields suitable for functional and structural characterization, as indicated by preliminary analyses using CD and NMR spectroscopy. Since the major hurdle in IDPs' production is their propensity to proteolytic degradation both in vivo and in vitro, it is important to identify expression and purification conditions that are compatible with the integrity of the final products. The Sic1 fragments prepared in this study will be subjected to further studies, to describe the structural properties of these protein domains in the absence of interactors.

The above reported analysis of $Sic1^{C\text{-}ter}$ fragments shows that the sequence extension toward the N-terminus, beyond residue 186, results in N-terminal trimming by proteolytic degradation, suggesting that the structural boundary of the C-terminal domain maps close to such position. As also indicated by other studies [40, 41], combined biochemical and bioinformatics analysis can lead to the identification of structural domains, even for IDPs. In the case of Sic1, the highly disordered N-terminal domain is the target of multiple phosphorylation events that trigger its degradative regulation [16, 17, 23]. The high flexibility of this Sic1 domain, as that of the corresponding region in the related mammalian protein p27, has been proposed to be essential in the mechanism of signal transduction [23, 42]. The more compact C-terminal domain mediates the interactions with Cdk1 and Clb5/6. Combined biochemical studies and molecular-dynamics simulations might help develop a structural model for this KID domain and decipher the role of its intrinsic structure in molecular recognition.

Acknowledgments

We thank Henriette Molinari for competent assistance acquiring and interpreting NMR spectra and Lilia Alberghina for helpful discussions. This study was supported by grants MIUR-Italbionet to Lilia Alberghina and grants from the University Milano-Bicocca ("Fondo Ateneo per la Ricerca") to R.G.

References

1. Fink, A. L. (2005). Natively unfolded proteins. Current Opinion in Structural Biology, 15, 35–41.

2. Uversky, V. N., Gillespie, J. R., & Fink, A. L. (2000). Why are ‘‘natively unfolded’’ proteins unstructured under physiologic conditions? Proteins, 41, 415–427.
3. Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., Obradovic, Z., et al. (2004). The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Research, 32, 1037–1049.
4. Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., et al. (2001). Intrinsically disordered protein. Journal of Molecular Graphics and Modelling, 19, 26–59.
5. Vacic, V., Oldfield, C. J., Mohan, A., Radivojac, P., Cortese, M. S., Uversky, V. N., et al. (2007). Characterization of molecular recognition features, MoRFs, and their binding partners. Journal of Proteome Research, 6, 2351–2366.
6. Uversky, V. N., Oldfield, C. J., & Dunker, A. K. (2005). Showing your ID: Intrinsic disorder as an ID for recognition, regulation and cell signaling. Journal of Molecular Recognition, 18, 343–384.
7. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., & Obradovic, Z. (2002). Intrinsic disorder and protein function. Biochemistry, 41, 6573–6582.
8. Dunker, A. K., & Obradovic, Z. (2001). The protein trinity– linking function and disorder. Nature Biotechnology, 19, 805–806.
9. Dyson, H. J., & Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. Nature Reviews Molecular Cell Biology, 6, 197–208.
10. Wright, P. E., & Dyson, H. J. (1999). Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. Journal of Molecular Biology, 293, 321–331.
11. Dyson, H. J., & Wright, P. E. (2002). Coupling of folding and binding for unstructured proteins. Current Opinion in Structural Biology, 12, 54–60.
12. Uversky, V. N., Oldfield, C. J., Midic, U., Xie, H., Xue, B., Vucetic, S., et al. (2009). Unfoldomics of human diseases: Linking protein intrinsic disorder with diseases. BMC Genomics, 10(1), S7.
13. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z., & Dunker, A. K. (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. Journal of Molecular Biology, 323,573–584.
14. Schwob, E., Bohm, T., Mendenhall, M., & Nasmyth, K. (1994). The B-type cyclin kinase inhibitor p40SIC1 controls the G1 to S transition in S. cerevisiae. Cell, 79, 233–244.
15. Mendenhall, M. D., & Hodge, A. E. (1998). Regulation of Cdc28 cyclin-dependent protein kinase activity during the cell cycle of the yeast Saccharomyces cerevisiae. Microbiology and Molecular Biology Reviews, 62, 1191–1243.
16. Verma, R., Annan, R. S., Huddleston, M. J., Carr, S. A., Reynard, G., & Deshaies, R. J. (1997). Phosphorylation of Sic1p by G1 Cdk required for its degradation and entry into S phase. Science, 278, 455–460.
17. Deshaies, R. J., & Ferrell, J. E., Jr. (2001). Multisite phosphorylation and the countdown to S phase. Cell, 107, 819–822.

18. Hodge, A., & Mendenhall, M. (1999). The cyclin-dependent kinase inhibitory domain of the yeast Sic1 protein is contained within the C-terminal 70 amino acids. Molecular and General
Genetics, 262, 55–64.
19. Borg, M., Mittag, T., Pawson, T., Tyers, M., Forman-Kay, J. D., & Chan, H. S. (2007). Polyelectrostatic interactions of disordered ligands suggest a physical basis for ultrasensitivity. Proceedings of the National Academy of Sciences of the United States of America, 104, 9650–9655.
20. Klein, P., Pawson, T., & Tyers, M. (2003). Mathematical modeling suggests cooperative interactions between a disordered polyvalent ligand and a single receptor site. Current Biology, 13,
1669–1678.
21. Escote´, X., Zapater, M., Clotet, J., & Posas, F. (2004). Hog1 mediates cell-cycle arrest in G1 phase by the dual targeting of Sic1. Nature Cell Biology, 10, 997–1002.
22. Brocca, S., Samalikova, M., Uversky, V. N., Lotti, M., Vanoni, M., Alberghina, L., et al. (2009). Order propensity of an intrinsically disordered protein, the cyclin-dependent-kinase inhibitor Sic1. Proteins, 76, 731–746.
23. Mittag, T., Orlicky, S., Choy, W. Y., Tang, X., Lin, H., Sicheri, F., et al. (2008). Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. Proceedings of the
National Academy of Sciences of the United States of America, 105, 17772–17777.
24. Sambrook, J., Fritsch, E. F., & Maniatis, T. (1989). Molecular cloning. A laboratory manual. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
25. Matsumura, I., & Rowe, L. A. (2005). Whole plasmid mutagenic PCR for directed protein evolution. Order propensity of a disordered protein, Sic1. Biomolecular Engineering, 22, 73–79.
26. Hanahan, D. (1985). Techniques for transformation of E. coli. In D. M. Glover (Ed.), DNA cloning: A practical approach (p. 109). Mc Lean, VA: IRL press.
27. Promega Subcloning Notebook 44. http://www.promega.com./guides/subcloning_guide/.
28. The Qiaexpressionist (June 2003). http://qiagen.com/literature/default.
29. Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature, 227, 680–685.
30. Sreerama, N., & Woody, R. W. (2000). Estimation of protein secondary structure from CD spectra: Comparison of CONTIN, SELCON and CDSSTR methods with an expanded reference set.
Analytical Biochemistry, 282, 252–260.
31. Hodge, A., & Mendenhall, M. (1999). The cyclin-dependent kinase inhibitory domain of the yeast Sic1 protein is contained within the C-terminal 70 amino acids. Molecular and General
Genetics, 1, 55–64.
32. Tompa, P. (2002). Intrinsically unstructured proteins. Trends in Biochemical Sciences, 27, 527–533.
33. Cortese, M. S., Baird, J. P., Uversky, V. N., & Dunker, A. K. (2005). Uncovering the unfoldome: Enriching cell extracts for unstructured proteins by acids treatment. Journal of Proteome Research, 4, 1610–1618.

34. Csizmok, V., Szollosi, E., Friedriich, P., & Tompa, P. (2006). A novel two-dimensional electrophoresis technique for the identification of intrinsically unstructured proteins. Molecular and Cellular Proteomics, 5, 265–273.
35. Galea, C. A., Pagala, V. R., Obenauer, J. C., Park, C. G., Slaughter, C. A., & Kriwacki, R. W. (2006). Proteomic studies of the intrinsically unstructured mammalian proteome. Journal of Proteome Research, 5, 2839–2848.
36. Frottin, F., Martinez, A., Peynot, P., Mitra, S., Holz, R. C., Giglione, C., et al. (2006). The proteomics of N-terminal methionine cleavage. Molecular and Cellular Proteomics, 5, 2336–2349.
37. Woody, R. W. (2009). Circular dichroism of intrinsically disordered proteins. In V. N. Uversky, S. Longhi (Eds.), Instrumental analysis of intrinsically disordered proteins: Assessing structure and conformation (pp. 303–321). New York, NY: John Wiley and Sons Ltd.
38. Mulder, F. A. A., Lundqvist, M., & Scheek, R. (2009). Nuclear magnetic resonance spectroscopy applied to (intrinsically) disordered proteins. In V. N. Uversky, S. Longhi (Eds.), Instrumental analysis of intrinsically disordered proteins: Assessing structure and conformation (pp. 61–87). New York, NY: John Wiley and Sons Ltd.
39. Crimmins, D. L., & Kao, J. L. (2007). The human cardiac hormone fragment N-terminal pro B-type natriuretic peptide is an intrinsically unstructured protein. Archives of Biochemistry and Biophysics, 461, 242–246.
40. He´brard, E., Bessin, Y., Michon, T., Longhi, S., Uversky, V. N., Delalande, F., et al. (2009). Intrinsic disorder in viral proteins genome-linked: Experimental and predictive analyses. Virology Journal, 16, 6–23.
41. Tompa, P., Fuxreiter, M., Oldfield, C. J., Simon, I., Dunker, A. K., & Uversky, V. N. (2009). Close encounters of the third kind: disordered domains and the interactions of proteins. Bioessays, 31, 328–335.
42. Galea, C. A., Nourse, A., Wang, Y., Sivakolundu, S. G., Heller, W. T., & Kriwacki, R. W. (2008). Role of intrinsic flexibility in signal transduction mediated by the cell cycle regulator, p27Kip1. Journal of Molecular Biology, 376(3), 827–838.

Electrospray ionization-mass spectrometry conformational analysis of isolated domains of an intrinsically disordered protein

Biotechnology Journal. 2011 Jan;6(1):96-100

Lorenzo Testa, Stefania Brocca, Maria Šamalikova, Carlo Santambrogio, Lilia Alberghina and Rita Grandori

Department of Biotechnology and Bioscience, University of Milano-Bicocca, Milan, Italy

Correspondence: Dr. Rita Grandori, Department of Biotechnology and Bioscience, University of Milano-Bicocca, Piazza della Scienza 2, 20126 Milan, Italy
E-mail: rita.grandori@unimib.it
Fax: +39-02-64483565

ABSTRACT

The highly dynamic and heterogeneous molecular ensembles characterizing intrinsically disordered proteins (IDP) in solution pose major challenges to the conventional methods for structural analysis. Electrospray ionization-mass spectrometry (ESI-MS) allows direct detection of distinct conformational components, effectively capturing also partially folded and short-lived states. We report the description of two complementary fragments (1–186 and 187–284) of the IDP Sic1, a cyclin-dependent protein kinase inhibitor of yeast Saccharomyces cerevisiae. Structural heterogeneity is noted in both cases, but the two fragments reveal slightly different conformational properties. The results are consistent with previously reported differences between the two protein moieties and corroborate the feasibility of IDP conformational analysis by ESI-MS.

Keywords

Charge state distributions · Cyclin-dependent protein kinase inhibitors · Nano-electrospray · Partially folded states · Sic1

Abbreviations: CSD, charge state distribution; **ESI**, electrospray ionization; **IDP**, intrinsically disordered protein

The fast growing fold class of intrinsically disordered proteins (IDP) (http://scop.mrc-lmb.cam.ac.uk/scop/) groups proteins that lack an ordered three-dimensional structure unless they bind to specific interaction partners [1]. In their free state, IDP populate a highly dynamic and heterogeneous ensemble characterized by a small content of secondary structure and variable degrees of structural compactness [2]. IDP typically perform important regulatory functions, such as cell cycle control and transcription regulation, and induce pleiotropic effects by binding to a high number of different physiological partners [2]. Thus, IDP represent an interesting protein class, both in terms of biological relevance and folding mechanisms.

Similar to chemically denatured proteins, IDP have a highly dynamic nature, thus posing major challenges to the techniques employed for their structural characterization. Important advances have been achieved in the application of biophysical methods to conformational analysis of IDP in their free state [3].A significant contribution can be given to this field by those techniques that directly identify the distinct conformational components of molecular ensembles rather than measuring average properties of a protein sample. Mass spectrometry (MS) offers such a possibility, thanks to the ion-sorting mechanism inherent to analyte detection [4]. Furthermore, due to conformational effects on the ionization process, protein charge-state distributions (CSD) obtained by non-denaturing electrospray ionization (ESI)-MS reveal different conformers coexisting in the original sample. Compact conformations acquire less charges than disordered ones, resulting in multimodal CSD for structurally heterogeneous states.Although ESI-MS detects ions in the gas phase, the proton transfer reactions that are sensitive to protein conformation take place in the condensed state, specifically labeling the different conformers present at the moment of transfer from the liquid to the gas phase [5].Thus, ESI-MS, particularly under the very mild ionization/desolvation conditions of nano-ESI-MS, has become a widely used tool for protein conformational analysis [5].

The Sic1 protein from *Saccharomyces cerevisiae* is a cyclin-dependent protein kinase inhibitor playing a main regulatory role in the G1-to-S phase transition [6]. Sic1 has been recognized to be intrinsically disordered in its whole length and to populate a collapsed state with little amount of secondary structure and considerable content of tertiary structure [7]. Structural, functional, and evolutionary characterization of this 284-amino acid protein has revealed a modular organization, suggesting that it is possible to recognize something equivalent to structural domains even in IDP [8]. Another example of structural modularity within an IDP is offered by α-synuclein [9].The C-terminal region of Sic1 contains the kinase

inhibitory domain [10], related to the well-known N-terminal inhibitory domain of the mammalian tumor suppressor proteins p21 and p27 [11, 12]. Limited-proteolysis patterns of Sic1 suggest that the Sic1 C-terminal region is relatively more compact than the N-terminus, with the boundary of the C-terminal domain lying close to position 186 [7, 8]. In this report, the two complementary Sic1 fragments 1–186 and 187–284, produced as recombinant His_6-tagged proteins [8], were subjected to conformational analysis by nano-ESI-MS.

The raw spectra of the isolated fragments under non-denaturing conditions are shown in Fig. 1A, C. The CSD are bimodal, indicating that the original samples are characterized by the coexistence of distinct conformers at the moment of desolvation [5, 13]. A broad peak envelope at relatively high charge states characterizes each spectrum, indicating the presence of highly disordered species. This component is centered around the 22+ ion for the N-terminal fragment and the 14+ ion for the C-terminal fragment. At much higher m/z values, a second component is detectable, with main charge state 11+ for the N-terminal fragment and 8+ for the C-terminal fragment. As better evident in the latter case, this peak envelope is also quite narrow, as typical of highly homogeneous conformational states. Thus, both fragments seem to transiently populate ordered, compact states. The low-charge component is a minor fraction of the population, with higher relative intensity in the spectrum of the C-terminal fragment. In agreement with its higher proteolysis resistance and lower scores by the predictor of naturally disordered regions (PONDR) [7, 14], these findings indicate a stronger intrinsic propensity of the C-terminal domain for ordered structures. Although His tags seem to affect the compactness of IDP [15], the here observed differences are likely caused by the different amino acid sequences of the two fragments since both the recombinant products contain a C-terminal His_6 tag.

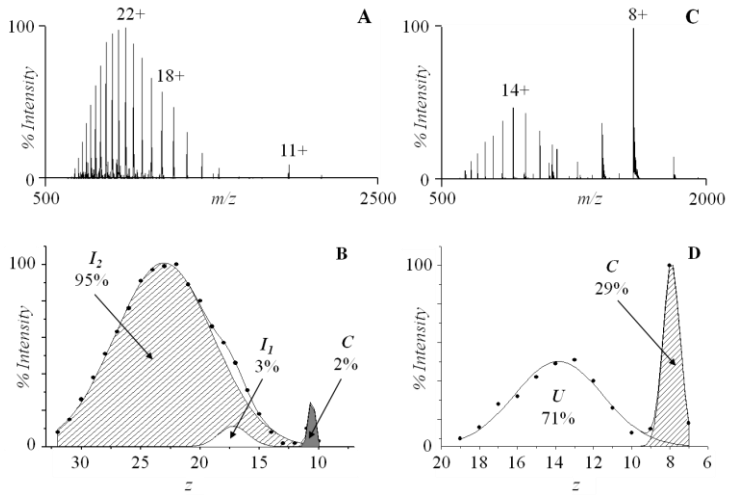


Figure 1. Conformational properties of Sic1$^{1–186}$ (A, B) and Sic1$^{187–284}$ (C, D) analyzed by nano-ESI-MS (A, C) and CSD deconvolution (B, D). Samples of 20 μM Sic1$^{1–186}$ or 10 μM Sic1$^{187–284}$ in 50 mM ammonium acetate, pH 6.5, were infused with the following instrumental setting: accumulation time 1 s, ion spray voltage 1200–1400 V, declustering potential 60–80 V, and instrument interface at room temperature. Gaussian fitting was performed by the software Origin7 (Originlab, Northampton, MA, USA), employing a series of Gaussian functions whose number and position were let vary until R^2 >0.99. Labels in (A) and (C) show the main charge state of each peak envelope. Labels in (B) and (D) indicate the identified conformers and their relative amounts. U, Unfolded; I, intermediate; C, compact form, as derived by CSD analysis (see text).

The coexisting different conformers can be better identified and quantified by Gaussian fitting of the CSD, after transformation from $x = m/z$ to $x = z$ as abscissa axis [16]. The results of such a deconvolution are shown in Fig. 1B, D. Although the signal yield of a given protein might be affected by its fold state [17], the areas derived by the fitting give an approximate indication of the relative amounts of the different species, particularly for comparative studies as in this case. The putative compact form turns out to represent only 2% of the population for the N-terminal fragment, while it is estimated around 29% for the C-terminal fragment. Furthermore, while the high-charge component can be fit by a single Gaussian function for the C-terminal fragment, two non-resolved species are identified in the asymmetric envelope of the N-terminal fragment. Nevertheless, the component centered on the 22+ ion is the largely predominant one (95%).

In order to achieve more confident identification of the distinct components detected above, conformational transitions were induced by the addition of

denaturants compatible with electrospray, such as acids and organic solvents. As can be seen in Fig. 2A, D, the addition of 1% formic acid results in complete loss of the low-charge component of the N-terminal fragment and in its strong decrease for the C-terminal fragment. The latter also disappears upon further addition of 50% acetonitrile (Fig. 2E). These results indicate that these species can be “denatured” by acids and organic solvents into more extended conformations. This observation also provides additional evidence that the low-charge components reflect structural properties of the analyzed proteins.

The disordered component of the C-terminal domain does not undergo major changes in response to pH and solvent hydrophobicity, with only a shift of its main charge state from 14+ to 15+ (Fig. 2D, E). Therefore, this component represents already an almost completely unfolded state even in the absence of denaturants. Conversely, the highly charged component of the N-terminal fragment shifts from main charge state 22+ to main charge state 27+, indicating that the maximally extended conformation is reached only by chemical denaturation (Fig. 2A, B).The molecular population in the presence of the cosolvents is still quite heterogeneous for the N-terminal domain, as indicated by a highly asymmetric CSD. Spectrum deconvolution by Gaussian fitting indicates persistence of two partially folded forms (20 and 10%), along with the major component (70%) of the fully unfolded protein (Fig. 2C).These intermediate species coincide with those detected in the absence of denaturants and are identified here as I1 and I2. Instead, a single component is detected for the C-terminal fragment in the presence of the denaturants (Fig. 2F).

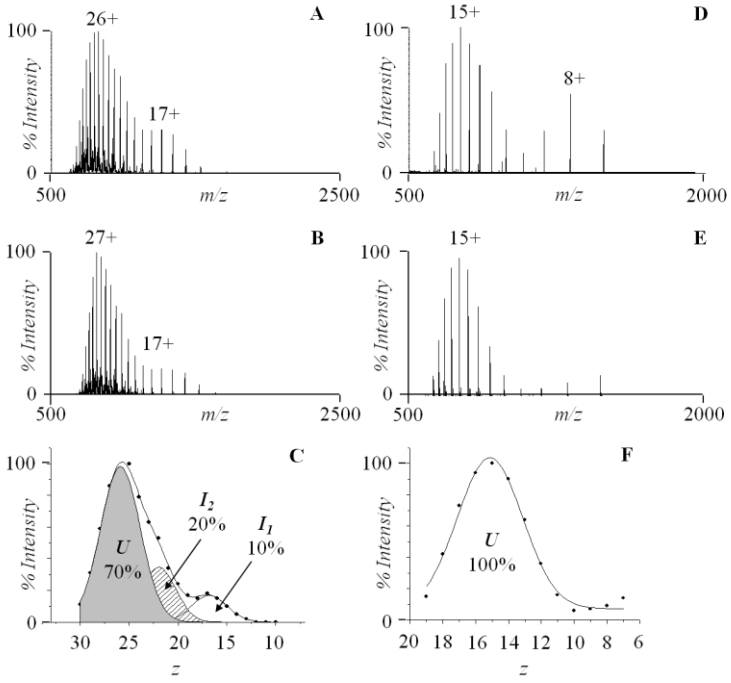


Figure 2. Conformational transitions of Sic1$^{1–186}$ (A–C) and Sic1$^{187–284}$ (D–F) monitored by nano-ESI-MS (A, B, D, E) and CSD deconvolution (C, F). Samples of 20 μM Sic1$^{1–186}$ or 10 μM Sic1$^{187–284}$ in 50 mM ammonium acetate, 1% formic acid (pH 3) (A, D) and 50 mM ammonium acetate, 1% formic acid, 50% acetonitrile (B, E) were infused with the same instrumental settings as in Fig. 1. Gaussian fitting was performed as in Fig. 1. The reported deconvolution data refer to the spectra in (B) and (E). Labels in (A), (B), (D) and (E) show the main charge state of each peak envelope. Labels in (C) and (F) indicate the identified conformers and their relative amounts. U, Unfolded; I, intermediate.

It is interesting to compare the here described spectra with the spectra of the full-length Sic1 protein that has been reported previously [7]. Under the same non-denaturing conditions, the full-length protein displays a single, broad CSD corresponding to a collapsed state of moderate compactness.

Therefore, while the CSD of the C-terminal fragment is sharply bimodal, this feature is barely detectable for the N-terminal fragment and not present at all for the full-length protein. These results suggest that highly compact conformations can arise only transiently and only locally.

The degree of compactness of the detected conformers can be evaluated by comparison with an empirical reference curve of average charge versus protein mass determined for native globular proteins [18].The experimental points follow a power law function of the molecular weight [18].The point

representing the full-length Sic1 protein, based on previously published data [7], is an outlier on such a graph, with an experimental average charge of 28.6+ and a calculated value of 15+ (Supporting Information Fig. S1). This result is consistent with a Sic1 conformational state that resembles more a collapsed chain than a folded globular protein. On the opposite, the putative compact forms of the isolated Sic1 domains approach the expected limit charge for normally folded globular proteins of the same size (13+ for the N-terminal fragment and 10+ for the C-terminal fragment).Therefore, such conformers seem to be characterized by considerable compactness. As indicated by the data reported above, these components represent only minor fractions of the molecular population and, therefore, likely represent metastable states. Metastable, compact states have been detected for other IDP [19] and for other Sic1-derived fragments [8, 20]. Experiments are under way in our laboratory to confirm the existence of such compact forms by other techniques.

Another striking difference among the CSD of the here considered proteins is that the most highly charged, less compact, components shift to different extents upon denaturation. The effect grows disproportionally with chain length, going from the C-terminal fragment (from 14+ to 15+) to the N-terminal fragment (from 22+ to 27+) to the full-length protein (from 28+ to 43+). These results are in agreement with the described effects of chain length on the hydrodynamic radius of folded, chemically denatured, and intrinsically disordered proteins [15]. Therefore, conformational analysis by ESI-MS seems to capture essential structural features of IDP, in comparison to chemically denatured and natively folded proteins.

This work was supported by grants MIUR-Italbionet to L.A. and grants from the University Milano-Bicocca ("Fondo Ateneo per la Ricerca") to R.G. and S.B.

The authors have declared no conflict of interest.

References

[1] Fink, A. L., Natively unfolded proteins. Curr. Opin. Struct. Biol. 2005, 15, 35–41.

[2] Uversky,V.N., Oldfield, C. J.,Dunker,A. K., Showing your ID: Intrinsic disorder as an ID for recognition, regulation and cell signaling. J. Mol. Recognit. 2005, 18, 343–384.

[3] Eliezer, D., Biophysical characterization of intrinsically disordered proteins. Curr. Opin. Struct. Biol. 2009, 19, 23–30.

[4] Konermann, L., Rosell, F. I., Mauk,A.G., Douglas, D. J., Acidinduced denaturation of myoglobin studied by time-resolved electrospray ionization mass spectrometry. Biochemistry
1997, 36, 6448–6454.
[5] Kaltashov, I. A.,Abzalimov, R. R., Do ionic charges in ESI MS provide useful information on macromolecular structure? J. Am. Soc. Mass Spectrom. 2008, 19, 1239–1246.
[6] Schwob, E., Bohm, T., Mendenhall, M. D., Nasmyth, K., The B-type cyclin kinase inhibitor p40SIC1 controls the G1 to S transition in S. cerevisiae. Cell 1994, 79, 233–244.
[7] Brocca, S., Samalikova, M., Uversky,V. N., Lotti, M. et al., Order propensity of an intrinsically disordered protein, the cyclin-dependent-kinase inhibitor Sic1. Proteins 2009, 76,
731–746.
[8] Brocca, S., Testa, L., Samalikova, M., Grandori, R. et al., Defining structural domains of an intrinsically disordered protein: Sic1, the cyclin-dependent kinase inhibitor of Saccharomyces
cerevisiae. Mol. Biotechnol. 2010, DOI: 10.1007/s12033-010-9309-y.
[9] Uversky,V.N.,Fink,A. L.,Amino acid determinants of alpha-synuclein aggregation: Putting together pieces of the puzzle. FEBS Lett. 2002, 522, 9–13.
[10] Hodge,A., Mendenhall, M.,The cyclin-dependent kinase inhibitory domain of the yeast Sic1 protein is contained within the C-terminal 70 amino acids. Mol. Gen. Genet. 1999, 262, 55–64.
[11] Barberis, M., De Gioia, L., Ruzzene, M., Sarno, S. et al., The yeast cyclin-dependent kinase inhibitor Sic1 and mammalian p27Kip1 are functional homologues with a structurally conserved inhibitory domain. Biochem. J. 2005, 387, 639–647.
[12] Galea, C. A., Wang, Y., Sivakolundu, S. G., Kriwacki, R. W., Regulation of cell division by intrinsically unstructured proteins: Intrinsic flexibility, modularity, and signaling conduits.
Biochemistry 2008, 47, 7598–7609.
[13] Borysik, A. J., Radford, S. E., Ashcroft, A. E., Co-populated conformational ensembles of beta2-microglobulin uncovered quantitatively by electrospray ionization mass spectrometry.
J. Biol. Chem. 2004, 279, 27069–27077.
[14] Peng, K., Vucetic, S., Radivojac, P., Brown, C. J. et al., Optimizing long intrinsic disorder predictors with protein evolutionary information. J. Bioinform. Comput. Biol. 2005, 3, 35–60.
[15] Marsh, J. A., Forman-Kay, J. D., Sequence determinants of compaction in intrinsically disordered proteins. Biophys. J. 2010, 98, 2383–2390.
[16] Dobo, A., Kaltashov, I. A., Detection of multiple protein conformational ensembles in solution via deconvolution of charge-state distributions in ESI MS. Anal. Chem. 2001, 73, 4763–4773.
[17] Kuprowski, M. C., Konermann, L., Signal response of coexisting protein conformers in electrospray mass spectrometry. Anal. Chem. 2007, 79, 2499–2506.
[18] Heck, A. J., Van Den Heuvel, R. H., Investigation of intact protein complexes by mass spectrometry. Mass Spectrom. Rev. 2004, 23, 368–389.

[19] Espinoza-Fonseca, L. M., Leucine-rich hydrophobic clusters promote folding of the N-terminus of the intrinsically disordered transactivation domain of p53. FEBS Lett. 2009, 583, 556–560.
[20] Mittag, T., Marsh, J., Grishaev, A., Orlicky, S. et al., Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase. Structure 2010, 18, 494–506.

Supporting Information

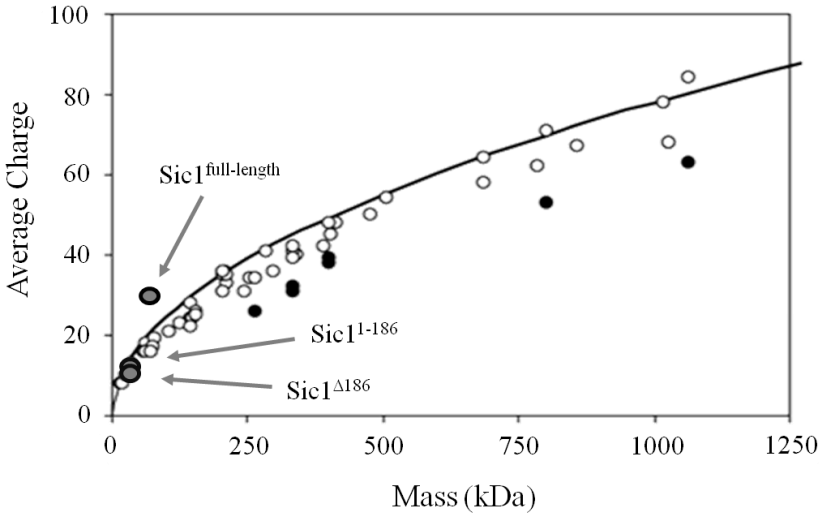


Figure S1. Evaluation of structural compactness of Sic1 and its fragments. The image depicts the charge-mass relation for a set of globular protein analyzed by ESI-MS, in positive (○) and negative (●) ion mode. The black line represents the Rayleigh-limit charge calculated for water droplets of the same size of the proteins. The experimental avarage charge for the full-length Sic1 (28.6 ±1.4) protein is derived from previously published data [7]. For the N-and C-terminal fragments, the values are derived from the low-charge components identified in this study (10.97 ±0.092 and 7.95 ±0.04, respectively). Each value is calculated as the average from at least five independent repetitions. Image modified with permission from [18].

Compaction Properties of an Intrinsically Disordered Protein: Sic1 and Its Kinase-Inhibitor Domain

Biophysical Journal. 2011 May 4; 100(9): 2243–2252.

Stefania Brocca,[†△] Lorenzo Testa,[†△] Frank Sobott,[‡] Maria Šamalikova,[†] Antonino Natalello,[†] Elena Papaleo,[†] Marina Lotti,[†] Luca De Gioia,[†] Silvia Maria Doglia,[†] Lilia Alberghina,[†] and Rita Grandori[†*]

† Department of Biotechnology and Biosciences, University of Milano-Bicocca, Milan, Italy
‡ Department of Chemistry, University of Antwerp, Antwerp, Belgium

△Stefania Brocca and Lorenzo Testa contributed equally to this work.

*Correspondence: rita.grandori@unimib.it

ABSTRACT

IDPs in their unbound state can transiently acquire secondary and tertiary structure. Describing such intrinsic structure is important to understand the transition between free and bound state, leading to supramolecular complexes with physiological interactors. IDP structure is highly dynamic and, therefore, difficult to study by conventional techniques. This work focuses on conformational analysis of the KID fragment of the Sic1 protein, an IDP with a key regulatory role in the
cell-cycle of Saccharomyces cerevisiae. FT-IR spectroscopy, ESI-MS, and IM measurements are used to capture dynamic and short-lived conformational states, probing both secondary and tertiary protein structure. The results indicate that the isolated Sic1 KID retains dynamic helical structure and populates collapsed states of different compactness. A metastable, highly compact species is detected. Comparison between the fragment and the full-length protein suggests that chain length is crucial to the stabilization of compact states of this IDP. The two proteins are compared by a length-independent compaction index.

Abbreviations used:

CKI, cyclin-dependent kinase inhibitor; KID, kinase inhibitory domain; IDP, intrinsically disordered protein; IFSU, intrinsically folded structural unit; NMR, nuclear magnetic resonance; FT-IR, Fourier-transform infrared; ESI, electrospray ionization; MS, mass spectrometry; IM, ion mobility; CSD, charge-state distribution; FSD, Fourier-self-deconvoluted; R_h, hydrodynamic radius; CI, compaction index.

INTRODUCTION

CKIs are key regulatory proteins that modulate kinase activities through the eukaryotic cell cycle. CKIs carry out their inhibitory function by the formation of ternary complexes with the target kinase and a cognate cyclin. The dynamic composition of kinase complexes throughout the cell cycle determines at which time and in which intracellular compartment regulatory phosphorylation events will take place (1).

Most CKIs belong to the class of IDPs (2,3). These proteins lack an ordered three-dimensional structure in their free state and undergo folding upon binding to specific interactors. Intrinsic structural disorder seems to favor multiplicity of interactions, independent regulation of affinity and specificity, regulation by post-translational modifications, and fast association kinetics (4). Nevertheless, no clear correlation emerges from large-scale surveys between degree of disorder and number of interactors (5). Structural characterization of IDPs in their free state is essential to the interpretation of their biochemical properties but it is technically challenging, due to their dynamic and heterogeneous conformation (6). The available evidence suggests that IDPs in the absence of interactors transiently populate partially ordered or compact states (7-13).

The protein Sic1 is a yeast CKI that binds to the target Cdk1 kinase in complex with the G1 cyclins Clb5,6 (14). The formation of these inhibitory complexes prevents phosphorylation of Cdk1 G1 substrates, which is required to enter S phase. The G1-to-S transition is executed only upon Sic1 degradation, which is, in turn, triggered by multiple Sic1 phosphorylations in its N-terminal region (10,15). Sic1 intracellular levels rise again in late M phase. Sic1 has also been implicated in other regulatory processes, such as mitosis exit (16), coordination between cell growth and cell cycle (17,18), and defense against osmotic stress (19). The Sic1 KID has been identified as the minimal protein fragment required for complementation of the ΔSic1

phenotype in vivo (20). Such a fragment corresponds to the C-terminal 70, out of 284, amino acids forming the native protein.

Previous work has shown that Sic1 is disordered in its whole length, with some intrinsic propensity to form ordered helical structure (7). The Sic1 molecule also displays remarkable compactness that makes it more similar to a collapsed chain than to a random coil (7,10). Such intrinsic tertiary structure can be denatured by acids and organic solvents (7). Fragments corresponding to the N- and C-terminal moieties have been produced, showing that the little content in secondary structure is distributed quite uniformly throughout the chain length, although the C-terminus is slightly more ordered than the N-terminus (21). More remarkable differences between the N- and C-terminal regions emerge from limited proteolysis experiments, which hint to the C-terminal ∼1/3 of the protein as the most resistant fragment (7). Consistent with such evidence, conformational analysis by ESI-MS suggests that the Sic1 C-terminal domain is more structured than its complementary N-terminal domain (22). Thus, altogether, functional and structural features point to a modular organization of this protein, despite its disordered nature. In vitro and in vivo degradation patterns suggest that the boundary of the more compact C-terminal region lies close to residue 186 (7,21).

The Sic1 C-terminal region is structurally and functionally related to the mammalian p21 and p27 KID that is located, instead, at the protein N-terminus (23). Detailed structural characterization of the p27 KID has been achieved by NMR and could be compared to the crystallographic structure of the ternary complex with cyclin A and Cdk2 (24,25). The residue-level description of secondary structure indicates that, in the absence of interactors, short chain segments transiently assume ordered backbone conformation. These secondary-structure elements resemble those observed in the bound state and correspond to docking elements of the interaction surfaces. Therefore, these IFSUs (4) represent on-pathway nuclei for complex formation. The mechanism of p27 binding and induced folding has been investigated by an array of different biophysical techniques (4). The ternary complex forms via a stepwise process where one end of the p27 KID first binds to cyclin A and the opposite end binds to Cdk2, covering the ATP-binding site. Thus, the inhibitor wraps on the surface of the cyclin A/Cdk2 complex in a highly extended conformation engaging its two ends in specific interactions with quite distant binding sites. As a result, it acts as a structural and thermodynamic staple in the ternary complex (26).

Despite the detailed characterization of p21 and p27 secondary structure, very little is known about KID tertiary structure for CKIs in the absence of

interactors. The goal of this work is to describe the conformational properties of Sic1 KID in its free state, both in terms of secondary and tertiary structure. The highly dynamic nature of IDPs makes it difficult to capture structural features that might arise only transiently in the unbound state. Here, conformational transitions of Sic1 KID are analyzed by complementary biophysical methods. Secondary structure is monitored by FT-IR spectroscopy. Highly dynamic helical elements are detected. Protein tertiary structure is probed by nondenaturing gel filtration, ESI-MS, and ESI-IM-MS. The extent of ionization by electrospray reflects protein compactness in the original solution, at the moment of transfer to the gas phase. The most compact states result in the lowest charge states (27-29). ESI-IM-MS resolves distinct conformers in the gas phase, based on their different shape, and allows for direct detection of compact species (30-32). The molecular ensemble of the isolated KID fragment is found to interconvert between collapsed states of different compactness. A small fraction of the population is found in a highly compact state. Comparison to the full-length protein hints to a critical role of chain length in determining the overall compaction of this IDP.

MATERIALS AND METHODS

Sequence analysis

The values of net charge per residues were derived according to Pappu and co-workers (33). Predictions of the R_h were performed according to Forman-Kay and co-workers (34). Sequence hydrophobicity was calculated by the Kyte and Doolittle approximation (35), using a window size of 5 amino acids by the ProtScale Expasy tool (http://expasy.org/tools/). The hydrophobicity of each residue was normalized to a scale between 0 and 1.

Protein expression and purification

The Escherichia coli strain BL21-Rosetta 2(DE3) (Novagen, EMD Chemicals, Darmstadt, DE) was used for heterologous expression of $Sic1^{\Delta 214}$ fused to a His_6 tag at its C-terminus (21). A glycine residue is inserted between the His_6 tag and the natural C-terminus of the protein (21). Cultures in low-salt Luria-Bertani medium were grown at 37°C and 220 rpm to an optical density at 600 nm (OD600) between 0.5 and 0.7. Cultures were then transferred to 30°C and induced with 200 μM isopropyl-β-D-thiogalactopyranoside. Cells were harvested after 2 h by centrifugation at

9400 × g at room temperature for 15 min and either immediately used for protein extraction or frozen at −20°C. Protein extraction and purification were performed as previously described (21). Protein concentration was assessed by the Bradford protein assay (Bio-Rad Laboratories, Hercules, CA).

FT-IR spectroscopy

The FT-IR absorption spectra of $Sic1^{\Delta 214}$ at a concentration of 800 μM in D_2O, 280 mM NaCl were collected in transmission using a Varian 670-IR spectrometer (Varian Australia Pty Ltd., Mulgrave VIC, AU), equipped with a nitrogen-cooled Mercury Cadmium Telluride detector, under accurate dry-air purging. A sample volume of 20 μL was placed in a temperature-controlled transmission cell with BaF_2 windows separated by a 150 μm teflon spacer. Spectra were collected under the following conditions: 2 cm^{-1} resolution, 25 kHz scan speed, 1000 scans coaddition, and triangular apodization. For temperature-variable experiments, the sample was heated from 10 to 100°C at a rate of 0.4°C/min.

The $Sic1^{\Delta 214}$ FT-IR spectra were obtained after subtraction of the solvent absorption, collected under the same conditions. Measured spectra were smoothed by a binomial function (11 points) and the second derivatives were obtained by the Savitsky–Golay method (third grade polynomial, 5 smoothing points). The FT-IR spectra in transmission of Candida rugosa Lipase 1 (CRL1) and α-synuclein in D_2O solutions were also measured at 2 cm^{-1} spectral resolutions and under conditions similar to the $Sic1^{\Delta 214}$ samples (13,36). Curve fitting of the Amide-I band as a linear combination of Gaussian components was performed as previously described (36,37). Briefly, the number of components and the initial values of their peak positions were taken from the second derivative and FSD spectra. The curve fitting was performed in two steps. First, peak wavenumbers were kept fixed and the other parameters were adjusted iteratively. In the second step, peak positions were allowed to change as well, to optimize the set of fitting functions. The final peak positions were found to be identical to those observed in the second derivative spectrum. The fractional area of each Gaussian function over the total area reflects the percentage of the corresponding secondary structure.

The above analysis was performed using the GRAMS/32 software (Galactic Industries Corporation, Salem, NH), with the exception of the FSD analysis, for which the Resolution-Pro software (Varian Australia Pty Ltd, Mulgrave VIC, AU) was employed.

Analytical gel filtration

Gel filtration was performed on an ÄKTA purifier liquid-chromatography system, using a prepacked, 30 cm × 1 cm SuperdexTM 75 HR column (GE Healthcare, Little Chalfont, UK). Chromatography was carried out at room temperature in 50 mM sodium phosphate, pH 8.0, 200 mM NaCl at a flow rate of 0.5 mL/min, and monitored by absorbance at 280 nm. A calibration curve was constructed using the following standards (0.75 mg/mL): bovine aprotinin (6,500 Da), horse cytochrome c (12,400 Da), horse myoglobin (17,600 Da), chicken ovalbumin (45,000 Da), equine apoferritin (80,000 Da) (Sigma Aldrich, St. Louis, MO), and recombinant green fluorescent protein (29,800 Da). A second calibration curve was constructed to calculate the R_h values of $Sic1^{FL}$ and $Sic1^{\Delta 214}$, by plotting the R_h values of the standards (ovalbumin, myoglobin, cytochrome c, and aprotinin) against their relatve elution volume (38-40).

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. The samples contained 10 μM $Sic1^{\Delta 214}$ in 50 mM ammonium acetate at the proper pH, which was adjusted by ammonium hydroxide or formic acid before protein addition. Metal-coated borosilicate capillaries (Proxeon, Odense, DK), with medium-length emitter tip of 1 μm internal diameter, were used for nanospray. 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, and 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 (22). These data were fitted by the minimal number of Gaussian functions leading to a stable fit. The relative peak areas were used to build the profile of the pH-dependent transition. Data from three independent acquisitions at each pH values were fitted by a Boltzmann sigmoidal function (final $R^2 > 0.98$). Fitting analyses were performed by the software OriginPro 7.5 (Originlab, Northampton, MA).

IM

IM-mass spectrometry was performed under nondenaturing conditions, by reconstituting samples into aqueous ammonium acetate buffer (pH 7.0) and removal of nonvolatile salts and buffers using micro bio-spin buffer exchange columns (Bio-Rad Laboratories). The protein sample (10 μM) was introduced into the mass spectrometer using nano-ESI with an Advion Triversa inlet system (Advion, Ithaca, NY). Mass-to-charge values and drift time of positive ions were determined using a Waters Synapt 1 T-wave IMS quadrupole-time-of-flight MS/MS system (Waters Corporation, Milford MA). Experimental parameters were chosen to optimize desolvation of sample ions, maintaining conditions that promote the transmission of intact protein ions, while possible disruption of native conformations is minimized (Sampling Cone 60 V, Extraction Cone 0 V, Trap Collision Energy 6.0 V, Transfer Collision Energy 4.0 V, Trap DC Bias 20.0 V, Backing pressure 1.47e0 mbar and Source Pressure 4.66e-4 mbar, IMS Wave Velocity 350 m/s, IMS Wave Height 8.0 V).

RESULTS AND DISCUSSION

Sequence analysis

Recent analyses of IDPs phase diagrams point to mean hydrophobicity and mean net charge as discriminating order parameters (33,34,41,42-43). The mean hydrophobicity value <H> for $Sic1^{\Delta 214}$ is quite low, 0.33, and close to the value calculated for the full-length Sic1 ($Sic1^{FL}$) protein (0.38, Table 1). Such a low hydrophobicity places Sic1 and its fragment in the region of the natively unfolded proteins in the bidimensional diagram developed by Uversky and co-workers (42). This result is consistent with the already reported conclusion that Sic1 is disordered in its whole length (7). To evaluate sequence polarity, the net charge per residue was calculated as the difference between the fraction of positively charged residues (f_+) and the fraction of negatively charged residues (f_-), according to Pappu and co-workers (33). The $f_+ - f_-$ values for $Sic1^{\Delta 214}$ and $Sic1^{FL}$ are well below the recognized threshold of 0.2, discriminating globular and coil-like IDPs. These results suggest that both $Sic1^{\Delta 214}$ and $Sic1^{FL}$ would preferentially populate collapsed states. In particular, according to Pappu and co-workers (33), these proteins would belong to the category of the weak polyelectrolytes/polyampholytes, characterized by relatively small values of

both parameters f_+ and f_-. Altogether, these sequence features would identify Sic1$^{\Delta 214}$ and Sic1^{FL} as disordered globules in the three-dimensional phase diagram proposed by Pappu and co-workers (33), where they would lie quite close to the axis origin.

	Sic1$^{\Delta 214}$		Sic1^{FL}	
	No tag	**Histidine tag**	**No tag**	**Histidine tag**
N	70	77	284	291
f_-	0.214	0.195	0.148	0.144
f_+	0.243	0.221	0.151	0.148
f_+-f_-	0.029	0.026	0.004	0.004
<H>	0.33	0.32	0.38	0.37
Q	2	2	1	1
P_{pro}	0.043	0.039	0.095	0.093
Calculated R_h (Å)	21.64 (20.27)	22.72 (19.07)	44.15 (43.79)	44.70 (39.84)
R_h by gel filtration (Å)	20.57 ± 0.22		31.25 ± 3.53	
CI by gel filtration	0.57 ± 0.03		0.77 ± 0.13	

TABLE 1. Disorder parameters for Sic1$^{\Delta 214}$ and Sic1^{FL}

N, sequence length. f_+ - f_-, net charge per residue as defined by Pappu and co-workers (33), where f_+ and f_- are the fractions of negatively and positively charged residues, respectively. <H>, mean Kyte-Doolittle hydropathy score (35). Q, difference between the total number of negatively charged and positively charged residues. P_{pro}, fractional proline content. R_h was calculated by the simple power-law model $R_h = R_0N^{\nu}$ (Eq. 2 of reference 34), with parameters for IDPs, R_0 = 2.49 Å and ν = 0.509, or (values in parentheses) by the sequence-based model. R_h = (AP_{pro} + B)(C|Q| + D)S_{his} x R_0N^{ν} (Eq. 6 of reference 34), where A = 1.24, B = 0.904, C = 0.00759, D = 0.963, S_{his} = 0.901 (correction factor for histidine tag), R_0 = 2.49 and ν = 0.509. The R_h values of folded and unfolded proteins used to calculate CI were derived from Eq. 2 of reference 34.

Secondary structure by FT-IR spectroscopy

The fragment produced in this work encompasses the sequence from residue 215 to the C-terminus and corresponds to the minimal Sic1 fragment retaining inhibitory activity (20,23). Protein expression and purification was carried out from a recombinant E. coli strain carrying plasmid pET21a[SIC1$^{\Delta 214}$] as previously described (21). The purified fragment was employed for secondary structure analysis by FT-IR spectroscopy in the Amide-I region (1600 to 1700 cm^{-1}). Data were acquired in D_2O, because random coil and α-helix contributions largely overlap in H_2O (44). The Sic1$^{\Delta 214}$ absorption spectrum measured at 10°C is reported in Fig. 1A. To resolve the Amide-I band constituents, FSD and second derivative were calculated from the raw spectrum. The results are reported in Fig. 1A, together with the Amide-I curve fitting into Gaussian components. The

second derivative spectrum is dominated by the ~1640 cm^{-1} band, which can be assigned to random-coil structure (44). In addition, other components were identified and their relative intensities quantified by Gaussian fitting from the ratio of their band area over the total Amide I area (36 ,44). They were found to be ~15% turns (1676 and 1670 cm^{-1}), ~30% α- and 3_{10}-helix (1656 cm^{-1}), ~40% random coil (1640 cm^{-1}), and ~15% β-strands (1627 cm^{-1}).

Comparison of $Sic1^{\Delta 214}$ second derivative spectrum with those of CRL1 (a natively folded α/β protein) and α-synuclein (a well-known IDP) highlights the disordered nature of $Sic1^{\Delta 214}$ (Supporting Material, Fig. S1). Indeed, its second derivative FT-IR spectrum is very similar to that of α-synuclein. Moreover, we should note that a protein containing ~30% of helical structure, such as CRL1, usually displays a strong and well-resolved peak in the second derivative spectrum (36). In the case of $Sic1^{\Delta 214}$, however, the contribution at 1656 cm^{-1} appears only as a shoulder of the main 1640 cm^{-1} peak, as expected for a component having a large band width (45). Indeed, in this case, the α-helical component is quite broad, with a band width comparable to that of the random-coil component. This result can be due to the presence of dynamic α-helical structures and possibly also to 3_{10} helices. It can be concluded that this protein fragment has a disordered nature, with a predominance of random coil and a certain intrinsic propensity to form ordered secondary structure elements.

FT-IR measurements were also performed at variable temperatures (Fig. 1, B and C). By heating the sample up to 100°C, the random-coil band around 1640 cm^{-1} in the second derivative spectra shifts to 1645 cm^{-1}, whereas the intensity of the 1656 cm^{-1}α-helical component seems to increase. As indicated by the profiles reported in Fig. 1C, these effects are small but reproducible. Furthermore, the thermal transition is reversible. The results suggest that a minor structural rearrangement is induced by heat, in agreement with previous data on the full-length protein (7).

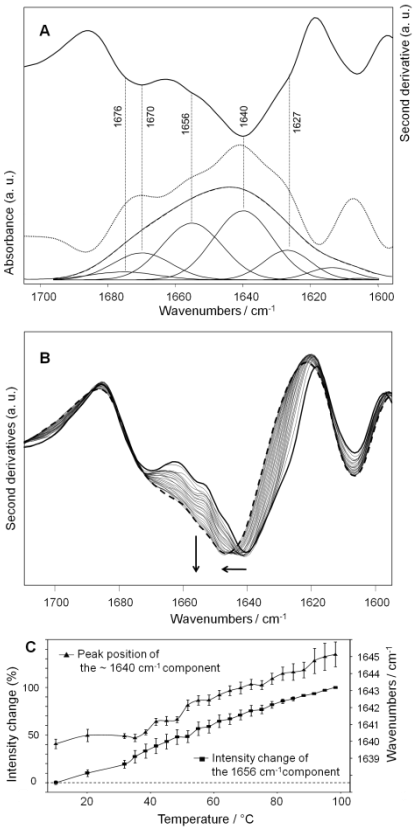


FIGURE 1 Sic1$^{\Delta 214}$ secondary structure by FT-IR spectroscopy. (A) Amide-I absorption spectrum (*continuous line*) and curve-fitted spectrum (*dashed line*) of Sic1$^{\Delta 214}$ in D_2O solution. The curve fitting of the spectrum as a linear combination of Gaussian components (*dot-dashed lines*) is superimposed to the measured spectrum. The Fourier self-deconvoluted (*dotted line*) and the second derivative (*continuous bold line*) spectra are also reported. (B) Amide-I second derivative spectra of Sic1$^{\Delta 214}$ in D_2O solution at different temperatures from 10°C (*bold continuous line*) to 100°C (*dashed line*). Arrows indicate the trend of spectral changes at increasing temperature. (C) Peak position of the ~1640 cm^{-1} component and intensity change of the 1656 cm^{-1} component as a function of temperature, both taken from the second derivative spectra.

Hydrodynamic behavior by gel filtration

The hydrodynamic behavior of the Sic1$^{\Delta 214}$ was characterized by gel filtration (Fig. 2). When standard globular proteins are used for column calibration, the elution time of Sic1$^{\Delta 214}$ yields an apparent molecular mass of 18,381 Da, well above the real value (9,293 Da, see below) (21). This result is consistent with a disordered nature of this protein. Based on the known R_h of the standards, the gel filtration data also allow estimating of the R_h value

of Sic1$^{\Delta 214}$ (40). The results indicate a value of 20.57 (± 0.22) Å. The same procedure has been applied, for comparison, to Sic1^{FL} yielding a value of 31.25 (± 3.53) Å (Fig. 2, Table 1). That these values actually refer to the intact products and that no proteolytic degradation had occurred during the analysis has been checked by SDS-electrophoresis of the eluted material.
It would be interesting to compare the degree of compaction of Sic1$^{\Delta 214}$ and Sic1^{FL} with reference to globular and denatured proteins. Relevant to this point, the R_h values of folded, globular proteins and those of chemically denatured proteins display a well-characterized dependence on the number of residues, N, and can be fit by simple power-law equations (34). Thus, we define the compaction index (*CI*):

$$CI = \frac{R_h^D - R_h}{R_h^D - R_h^F}$$

where R_h is the experimental value for a given protein and R_h^D and R_h^F are the reference values calculated for a chemically denatured or a globular, folded protein of the same size. Different from the frequently used parameter R_h/R_h^D, *CI* does not depend on *N* and, therefore, it allows for comparison among proteins of different lengths. The *CI* value can vary between 0 and 1, with 0 indicating minimal compaction and 1 maximal compaction. This index could be applied to any structural feature linked to compactness (also the gyration radius, for instance), given that the reference curves for folded and unfolded proteins used in the calculation are built to fit the same parameter. By using the reference equations for R_h published by Forman-Kay and co-workers (34), we obtain *CI* values of 0.57 (± 0.03) for Sic1$^{\Delta 214}$ and 0.77 (± 0.13) for Sic1^{FL}. Slightly different values are obtained using the equations published by Uversky (40), namely, 0.63 (± 0.02) for Sic1$^{\Delta 214}$ and 0.80 (± 0.13) for Sic1^{FL}. These results indicate that the full-length protein displays higher overall compaction than the fragment. Because the isolated N-terminal moiety is even more disordered than the here considered C-terminal fragment (22), it can be conceived that either long-range coupling between the N- and C- termini or indirect effects of chain length play a role in the compaction of the full-length protein. Nevertheless, gel filtration offers relatively low resolution and low sensitivity, and captures only average features of the molecular ensemble.

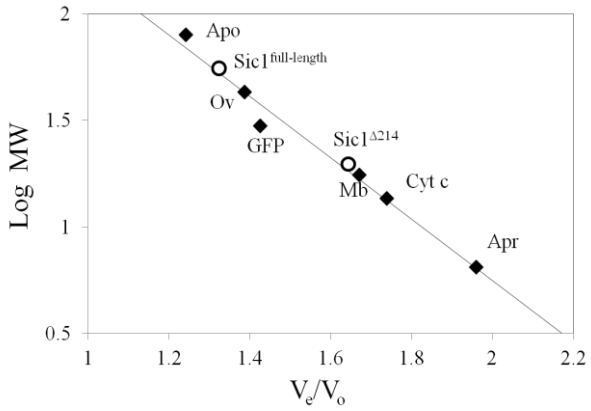


FIGURE 2 Hydrodynamic behavior of Sic1$^{\Delta 214}$ and Sic1^{FL} compared to standard globular proteins. Analytical gel filtration of ~100 μM protein in 50 mM sodium phosphate, pH 8.0, 200 mM NaCl. V_e, elution volume; V_o, void volume; Apo, apoferritin; Ov, ovalbumin; GFP, green fluorescent protein; Mb, myoglobin; Cyt c, cytochrome c; Apr, aprotinin. The reported values are the average from three independent experiments. Standard deviation was within 6.5%.

Tertiary structure by ESI-MS and ESI-IM-MS

Finer characterization of tertiary structural features of Sic1$^{\Delta 214}$ was achieved by ESI-MS. Protein CSDs obtained by ESI-MS are strongly affected by the global compactness of the protein structure at the moment of transfer from the liquid to gas phase (28,46). It has been shown that the ionization behavior can effectively distinguish not only folded and unfolded protein, but also partially folded forms from both the above (27-29). Furthermore, this kind of analysis offers the typical advantage of MS methods, namely to describe heterogeneous samples by direct detection of the distinct components (29). Thanks to these features, ESI-MS is particularly well suited to the identification of labile compact forms that might arise transiently within a molecular population. The spectra of Sic1$^{\Delta 214}$ under nondenaturing conditions are characterized by sharply bimodal distributions, with a predominant form centered on the 10+ ion and a minor component centered on the 7+ ion (Fig. 3A). Such a distribution is typical for coexistence of two distinct conformers of different compactness that do not interconvert on the time scale of the proton-transfer reactions that label protein molecules during the ESI process. The peak envelope at lower charge states (7+) would reflect the more compact form. Bimodal distributions have been observed on other Sic1-derived fragments (22). The relative contributions of the two components to the CSDs can be quantified

by Gaussian fitting, upon transformation of the spectra to $x = z$ abscissa axis (22). The results of such a procedure are reported in the insets of Fig. 3. Although quantitation of different conformers by ESI-MS can only be taken indicatively (47), the apparent relative amount of the compact form in 50 mM ammonium acetate, pH 6.5 is quite small (~20%). Mass deconvolution yields a value of 9293.5 Da (+/− 0.1 Da), consistent with the sequence of the fragment with the His_6 tag and without initial methionine (calculated mass 9293.4 Da).

To confirm that the two peak envelopes correspond to distinct conformational states, the effects of acids and organic solvents have been investigated (Fig. 3B and C). By the addition of 1% formic acid, the more compact form decreases and the highly charged component accumulates. By the further addition of 50% acetonitrile, the compact form disappears completely. These results suggest that the component centered on the 7+ ion has considerable tertiary structure that is destabilized by acids and organic solvents. The pH dependence reveals a quite cooperative transition (Fig. 3D). This response to pH might reflect the transition from disordered globule to swollen coil at increasing $f_+ - f_-$ in the phase diagram proposed by Pappu and co-workers (33). The pH response of this protein fragment is consistent with its amino acid sequence, counting 15 acidic residues (Fig. 3E). Although the His_6 tag does not change significantly the pI (9.13 with tag and 9.11 without tag) its presence might affect the shape of the transition. These effects of denaturing agents on the CSD provide important control evidence that the low-charge component in $Sic1^{\Delta 214}$ spectra is not artifactual. The high-charge component, instead, responds to denaturing agents only by a minor shift in the main charge state from 10+ to 11+. Therefore, the 11+ form can be identified as the most unfolded species, whereas the 10+ component is characterized by modest intrinsic compactness. We name the 10+ and 7+ components, respectively, C1 and C2, referring to collapsed species of different compactness.

Altogether, this evidence strongly suggests that the component centered on the 7+ ion reflects a metastable compact state of this molecule in solution. No significant changes were observed by varying the buffer concentration between 10 and 50 mM. It should be underscored that a net charge 7+ reflects a surprisingly high compactness for this protein fragment. Indeed, it closely matches the average charge (7.5) calculated for a globular, folded protein of this size, according to the empiric charge-to-mass relation in protein ESI-MS (48). The effect of pH has been investigated also by FT-IR spectroscopy. The second derivative spectrum at pH 1.5 displays a small reduction of the shoulder around 1627 cm , which characterizes the spectrum

at pH 6.5 (Fig. S2). This difference suggests that also minor rearrangements in protein secondary structure take place at acidic pH. The region influenced by these changes is that of β-strands.

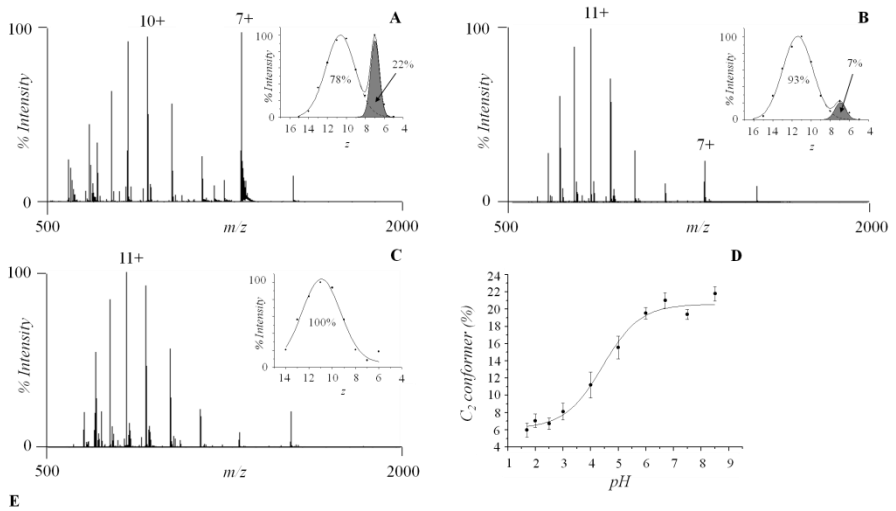


FIGURE 3 Sic1$^{\Delta 214}$ tertiary structure and unfolding. Nano-ESI-MS spectra of 10 μM protein in (A) 50 mM ammonium acetate, pH 6.5; (B) 50 mM ammonium acetate, 1% formic acid (pH ~3); and (C) 50 mM ammonium acetate, 1% formic acid, 50% acetonitrile. The main peak of each component is labeled by the corresponding charge state. The results of Gaussian fitting of the CSDs are shown in the insets. (D) Profile of the pH transition, showing the relative amount of the collapsed species C2 over the total population as a function of pH. The line represents data fitting by a sigmoidal curve. Error bars indicate the standard deviation from three independent acquisitions. (E) Amino acid sequence of Sic1$^{\Delta 214}$.

The existence of compact conformations of the here considered proteins was further tested by ESI-IM-MS measurements (Figure 4 and Figure 5). This method adds a second dimension to conventional ion sorting by ESI-MS, based on their mobility through a buffer gas under the influence of an electric field (30, 31 and 32). Under these conditions, IM is directly related to the shape and overall topology of the ion, leading to separation of the ions on the basis of their structural compactness. Although protein charging during electrospray is affected by the structural properties of the protein in solution, IM is, obviously, affected by the actual conformation held by the protein in the gas phase. Folded or partially folded conformations can survive upon desolvation for low charge states. Highly charged ions, instead, tend to open into more extended conformations due to electrostatic

repulsions, especially when they become activated due to collisions with background gas during transit of the mass spectrometer. The drift time plots obtained with Sic1$^{\Delta 214}$ (Fig. 4), under conditions that minimize thermal (collisional) activation of the ions, reveal two well-resolved populations. Because IM depends on the charge of the ions as well, particles with the same collisional cross sections come to lie on oblique lines (indicated in the figure). The species with the lowest mobility is enriched at high charge states and represents the least compact conformer. A second species with higher mobility is detectable and is, therefore, characterized by higher compactness. This species is maximally enriched at the charge states 5+ and 6+. The arrival time distribution for charge state 6+ is reported in Fig. 4B. These results provide direct evidence of the existence of compact conformations of Sic1$^{\Delta 214}$, clearly distinguishable from the extended form of the protein. Therefore, IM analysis supports the hypothesis that the observed bimodal CSD is due to coexistence of at least two distinct conformers.

The spectrum reported in Fig. 4C also reveals a small fraction of dimer-specific peaks at high m/z values, suggesting that Sic1$^{\Delta 214}$ might have an intrinsic propensity to dimerize. Dimers can be identified in the high-m/z range of the IM-plot (Fig. 4A). However, no dimers are detected for the same sample under the other instrumental setup employed for this work (Fig. 3). Therefore, Sic1 dimers must be highly unstable and prone to dissociate under electrospray conditions. More detailed studies will be needed to evaluate the structural and functional relevance of this dimeric species. It cannot be ruled out that they simply derive from low-level nonspecific association. The dimeric assemblies also reflect the existence of two conformational states of the protein subunits.

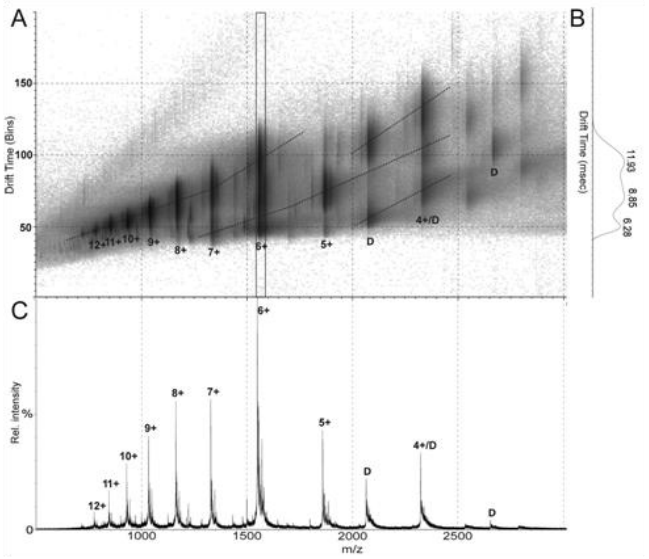


FIGURE 4 Analysis of Sic1$^{\Delta 214}$ distinct conformers in the gas phase with nondenaturing nano-ESI-IMS-MS. (A) IM-plot reporting the drift time measurements for each charge state. The dotted lines connect corresponding conformations that differ in IM due to their different charges. (B) Arrival time distribution for the 6+ charge state. (C) Corresponding nano-ESI-MS spectrum. The infused sample was 5 μM protein in 50 mM ammonium acetate pH 6.5.

The same analysis has been carried out on the intact protein, Sic1^{FL} (Fig. 5). In this case, the ESI-IM-MS 2D plots reveal three distinct populations (Fig. 5A). The predominant one displays the lowest mobility and corresponds, therefore, to the most extended conformation. As expected, this species is enriched at the highest charge states (34+ to 20+). A second population with intermediate mobility is detectable for the charge states between 20+ and 12+. Finally, a well-resolved series of peaks corresponding to the highest mobility is resolved in the low-charge region of the plot. This species is detectable for the charge states between 13+ and 9+ and is highly enriched for the 10+ ions. The arrival time distribution for this charge state is reported in Fig. 5B. These data indicate that, under nondenaturing conditions, collapsed states can be identified also for the intact protein. The analysis of Sic1^{FL} CSD (Fig. 5C) indicates that the molecular ensemble is more homogeneous than observed for Sic1$^{\Delta 214}$. Indeed the CSD is unimodal, rather than bimodal, with only some tailing to the right side. Thus, the large majority of Sic1^{FL} molecules are found in a compact state. That the component centered around the 30+ ion corresponds to a partially folded species in solution is also supported by previous denaturation studies (7).

Although the presence of histidine tags has been shown to slightly affect compactness of IDPs (34), the differences emerging from this comparison should be ascribed to the different primary structure of the two proteins, because both products contain a His6 tag.

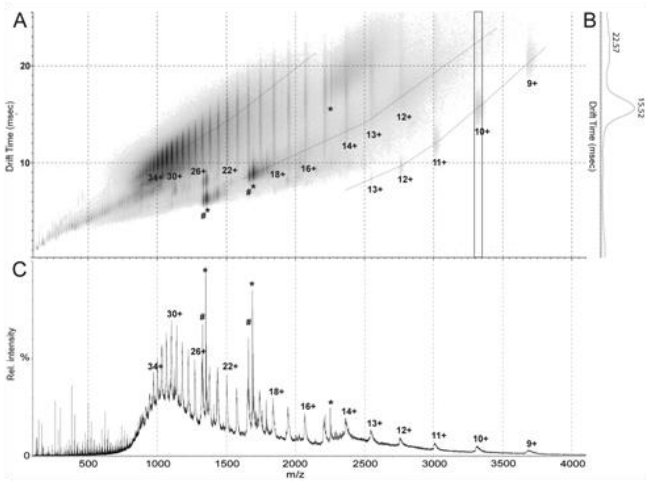


FIGURE 5 Analysis of Sic1^{FL} distinct conformers in the gas phase with nondenaturing nano-ESI-IMS-MS. (A) IM-plot reporting the drift time measurements for each charge state. The dotted lines connect corresponding conformations that differ in IM due to their different charges. (B) Arrival time distribution for the 10+ charge state. (C) Corresponding nano-ESI-MS spectrum. The infused sample was 5 μM protein in 50 mM ammonium acetate pH 6.5. The signals of proteolytic fragments are labeled with * and # in panels A and C.

Fig. S3 summarizes the results reported here, comparing calculated and experimental R_h values for both Sic1$^{\Delta 214}$ and Sic1^{FL}. The curve describing the dependence of R_h on chain length for IDPs in general (34) predicts the highest values. Slightly smaller values are calculated by the more accurate model that uses sequence-specific information (34). The average R_h measured by gel filtration for Sic1$^{\Delta 214}$ is very close to the calculated value. We attribute this R_h value to the highly predominant C1 component. Therefore, the C1 conformer of Sic1$^{\Delta 214}$ approximates the average compactness of IDPs for this chain length. The C2 conformer is not included in this plot because it could not be resolved by gel filtration, but would lie well below C1. The average experimental value for Sic1^{FL} is much below both theoretical values calculated for the intact protein. This result is consistent with MS data, indicating that most of the Sic1^{FL} protein is found in a highly collapsed state. Altogether, these results indicate that both Sic1^{FL}

and Sic1$^{\Delta 214}$ have a propensity to populate conformational states with higher compactness than predicted by the average behavior of IDPs.

Conclusion

Rapidly growing evidence indicates that IDPs in their unbound state (7, 8, 9, 10, 11, 12, 33 and 34) not only are far from resembling a true random coil, but are also quite different from chemically denatured proteins. However, the degree of compaction for IDPs is more variable than observed for folded, globular proteins (34). Sic1 is an example of IDP populating collapsed states of particularly high compactness. Nevertheless, compact states are destabilized in the isolated C-terminal fragment, relative to the native protein. Indeed, only a minor fraction of the molecular population is found in a compact state for Sic1$^{\Delta 214}$, consistently with the quite low CI value measured by gel filtration. The results reported here suggest that chain length could play an important role in stabilizing compact states of poorly foldable sequences.

The picture that emerges for the isolated Sic1 KID fragment is that of a highly dynamic structure, in which transient secondary and tertiary structure can develop. The detailed NMR studies performed on the related proteins p21 and p27 have shown that IFSUs are on-pathway for complex formation with the binding partners cyclin and kinase (4). However, highly compact tertiary structure might not be productive for complex formation, because the inhibitor seems to bind to its interactors in an extended conformation (23 and 24). More studies will be needed to understand whether the compact states detected here are on-pathway intermediates in the formation of Sic1 physiological complexes. Furthermore, it still remains to be clarified which interactions are responsible for chain compaction and which residues they involve.

The highly dynamic and heterogeneous molecular ensemble representing unfolded proteins poses a major challenge to structural characterization by biophysical methods. Short-lived structured states can, indeed, easily escape detection. Nevertheless, compact states of IDPs have been detected by several different biophysical techniques, such as small-angle x-ray scattering (10 and 49), atomic-force microscopy (50), Förster resonance energy transfer (51), ESI-MS (13), ESI-IM-MS (52), NMR studies on translational diffusion coefficients (10), and paramagnetic relaxation enhancement (10 and 12). The ion-sorting process inherent to MS measurements offers the possibility to directly assess species distributions in complex mixtures, allowing detection of metastable conformational states. Combining this information with

structural data derived from IM measurements can offer an in-depth description of highly dynamic molecular systems.

Acknowledgements

This work was supported by the grants FIRB/Italbionet to L.A., FAR ("Fondo Ateneo per la Ricerca") of the University Milano-Bicocca to R.G., M.L., S.B., and S.M.D., and by postdoctoral fellowships of the University of Milano-Bicocca to A.N. and M.Š.

REFERENCES

1. Alfieri, R., M. Barberis, ., L. Alberghina. 2009. Towards a systems biology approach to mammalian cell cycle: modeling the entrance into S phase of quiescent fibroblasts after serum stimulation. BMC Bioinformatics. 10 (Suppl 12):S16.
2. Fink, A. L. 2005. Natively unfolded proteins. Curr. Opin. Struct. Biol. 15:35–41.
3. 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. 102:73–84.
4. Galea, C. A., Y. Wang, ., R. W. Kriwacki. 2008. Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. Biochemistry. 47:7598–7609.
5. Schnell, S., S. Fortunato, and S. Roy. 2007. Is the intrinsic disorder of proteins the cause of the scale-free architecture of protein-protein interaction networks? Proteomics. 7:961–964.
6. Eliezer, D. 2009. Biophysical characterization of intrinsically disordered proteins. Curr. Opin. Struct. Biol. 19:23–30.
7. Brocca, S., M. Samalıkova´,., R. Grandori. 2009. Order propensity of an intrinsically disordered protein, the cyclin-dependent-kinase inhibitor Sic1. Proteins. 76:731–746.
8. Galea, C. A., A. Nourse, ., R. W. Kriwacki. 2008. Role of intrinsic flexibility in signal transduction mediated by the cell cycle regulator, p27 Kip1. J. Mol. Biol. 376:827–838.
9. Tran, H. T., A. Mao, and R. V. Pappu. 2008. Role of backbone-solvent interactions in determining conformational equilibria of intrinsically disordered proteins. J. Am. Chem. Soc. 130:7380–7392.
10. Mittag, T., J. Marsh, ., J. D. Forman-Kay. 2010. Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase. Structure. 18:494–506.
11. Ganguly, D., and J. Chen. 2009. Atomistic details of the disordered states of KID and pKID. Implications in coupled binding and folding. J. Am. Chem. Soc. 131:5214–5223.

12. Salmon, L., G. Nodet,., M. Blackledge. 2010. NMR characterization of long-range order in intrinsically disordered proteins. J. Am. Chem. Soc. 132:8407–8418.
13. Natalello, A., F. Benetti, ., R. Grandori. 2011. Compact conformations of a-synuclein induced by alcohols and copper. Proteins. 79:611–621.
14. Deshaies, R. J. 1997. Phosphorylation and proteolysis: partners in the regulation of cell division in budding yeast. Curr. Opin. Genet. Dev. 7:7–16.
15. Deshaies, R. J., and J. E. Ferrell, Jr. 2001. Multisite phosphorylation and the countdown to S phase. Cell. 107:819–822.
16. Lopez-Avile´s, S., O. Kapuy, ., F. Uhlmann. 2009. Irreversibility of mitotic exit is the consequence of systems-level feedback. Nature. 459:592–595.
17. Coccetti, P., R. L. Rossi,., L. Alberghina. 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.
18. Barberis, M., E. Klipp, ., L. Alberghina. 2007. Cell size at S phase initiation: an emergent property of the G1/S network. PLOS Comput. Biol. 3:e64.
19. Yaakov, G., A. Duch, ., F. Posas. 2009. The stress-activated protein kinase Hog1 mediates S phase delay in response to osmostress. Mol. Biol. Cell. 20:3572–3582.
20. 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. 262:55–64.
21. Brocca, S., L. Testa,., M. Lotti. 2011. Defining structural domains of an intrinsically disordered protein: Sic1, the cyclin-dependent kinase inhibitor of Saccharomyces cerevisiae. Mol. Biotechnol. 47:34–42.
22. Testa, L., S. Brocca, ., R. Grandori. 2011. Electrospray-ionization mass spectrometry conformational analysis of isolated domains of an intrinsically disordered protein. Biotechnol. J. 6:96–100.
23. Barberis, M., L. De Gioia, ., L. Alberghina. 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.
24. Russo, A. A., P. D. Jeffrey, ., N. P. Pavletich. 1996. Crystal structure of the p27Kip1 cyclin-dependent-kinase inhibitor bound to the cyclin A-Cdk2 complex. Nature. 382:325–331.
25. 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. 353:1118–1128.
26. Bowman, P., C. A. Galea, ., R. W. Kriwacki. 2006. Thermodynamic characterization of interactions between p27(Kip1) and activated and non-activated Cdk2: intrinsically unstructured proteins as thermodynamic tethers. Biochim. Biophys. Acta. 1764:182–189.
27. Samalikova, M., C. Santambrogio, and R. Grandori. 2010. Mass-spectometry tools for the investigation of structural disorder and conformational transitions in proteins. In Instrumental Analysis of Intrinsically Disordered Proteins. V. N. Uversky and S. Longhi, editors. John Wiley & Sons, Hoboken, NJ. 629–652.
28. Kaltashov, I. A., and R. R. Abzalimov. 2008. Do ionic charges in ESI MS provide useful information on macromolecular structure? J. Am. Soc. Mass Spectrom. 19:1239–1246.

29. Borysik, A. J., S. E. Radford, and A. E. Ashcroft. 2004. Co-populated conformational ensembles of beta2-microglobulin uncovered quantitatively by electrospray ionization mass spectrometry. J. Biol. Chem. 279:27069–27077.
30. Smith, D. P., K. Giles, ., A. E. Ashcroft. 2007. Monitoring copopulated conformational states during protein folding events using electrospray ionization-ion mobility spectrometry-mass spectrometry. J. Am. Soc. Mass Spectrom. 18:2180–2190.
31. Smith, D. P., T. W. Knapman, ., A. E. Ashcroft. 2009. Deciphering drift time measurements from travelling wave ion mobility spectrometry- mass spectrometry studies. Eur. J. Mass Spectrom. (Chichester, Eng.). 15:113–130.
32. Uetrecht, C., R. J. Rose, ., A. J. Heck. 2010. Ion mobility mass spectrometry of proteins and protein assemblies. Chem. Soc. Rev. 39:1633–1655.
33. Mao, A. H., S. L. Crick,., R. V. Pappu. 2010. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. Proc. Natl. Acad. Sci. USA. 107:8183–8188.
34. Marsh, J. A., and J. D. Forman-Kay. 2010. Sequence determinants of compaction in intrinsically disordered proteins. Biophys. J. 98:2383–2390.
35. Kyte, J., and R. F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157:105–132.
36. Natalello, A., D. Ami, ., S. M. Doglia. 2005. Secondary structure, conformational stability and glycosylation of a recombinant Candida rugosa lipase studied by Fourier-transform infrared spectroscopy. Biochem. J. 385:511–517.
37. Vila, R., I. Ponte,., P. Suau. 2001. Induction of secondary structure in a COOH-terminal peptide of histone H1 by interaction with the DNA: an infrared spectroscopy study. J. Biol. Chem. 276:30898–30903.
38. Irvine, G. B. 2001. Determination of molecular size by size-exclusion chromatography (gel filtration). Curr. Protoc. Cell. Biol., 5.5.
39. Mohr, D., S. Frey,., D. Go¨rlich. 2009. Characterisation of the passive permeability barrier of nuclear pore complexes. EMBO J. 28:2541–2553.
40. Uversky, V. N. 1993. Use of fast protein size-exclusion liquid chromatography to study the unfolding of proteins which denature through the molten globule. Biochemistry. 32:13288–13298.
41. Uversky, V. N. 2002. What does it mean to be natively unfolded? Eur. J. Biochem. 269:2–12.
42. Uversky, V. N., J. R. Gillespie, and A. L. Fink. 2000. Why are ''natively unfolded'' proteins unstructured under physiologic conditions? Proteins. 41:415–427.
43. Muller-Spath, S., A. Soranno, ., B. Schuler. 2010. From the Cover: Charge interactions can dominate the dimensions of intrinsically disordered proteins. Proc. Natl. Acad. Sci. USA. 107:14609–14614.
44. Arrondo, J. L. R., and F. M. Gon˜ i. 1999. Structure and dynamics of membrane proteins as studied by infrared spectroscopy. Prog. Biophys. Mol. Biol. 72:367–405.
45. Susi, H., and D. M. Byler. 1986. Resolution-enhanced Fourier transform infrared spectroscopy of enzymes. Methods Enzymol. 130: 290–311.
46. Ashcroft, A. E. 2010. Mass spectrometry and the amyloid problem—how far can we go in the gas phase? J. Am. Soc. Mass Spectrom. 21:1087–1096.

47. Kuprowski, M. C., and L. Konermann. 2007. Signal response of coexisting protein conformers in electrospray mass spectrometry. Anal. Chem. 79:2499–2506.
48. Heck, A. J., and R. H. Van Den Heuvel. 2004. Investigation of intact protein complexes by mass spectrometry. Mass Spectrom. Rev. 23:368–389.
49. Bernado´ , P., E. Mylonas, ., D. I. Svergun. 2007. Structural characterization of flexible proteins using small-angle X-ray scattering. J. Am. Chem. Soc. 129:5656–5664.
50. Brucale, M., M. Sandal, ., B. Samorı`. 2009. Pathogenic mutations shift the equilibria of alpha-synuclein single molecules towards structured conformers. ChemBioChem. 10:176–183.
51. Nettels, D., S. Mu¨ller-Spa¨th, ., B. Schuler. 2009. Single-molecule spectroscopy of the temperature-induced collapse of unfolded proteins. Proc. Natl. Acad. Sci. USA. 106:20740–20745.
52. Bernstein, S. L., D. Liu, ., J. R. Winkler. 2004. Alpha-synuclein: stable compact and extended monomeric structures and pH dependence of dimer formation. J. Am. Soc. Mass Spectrom. 15:1435–1443.

Supporting Material

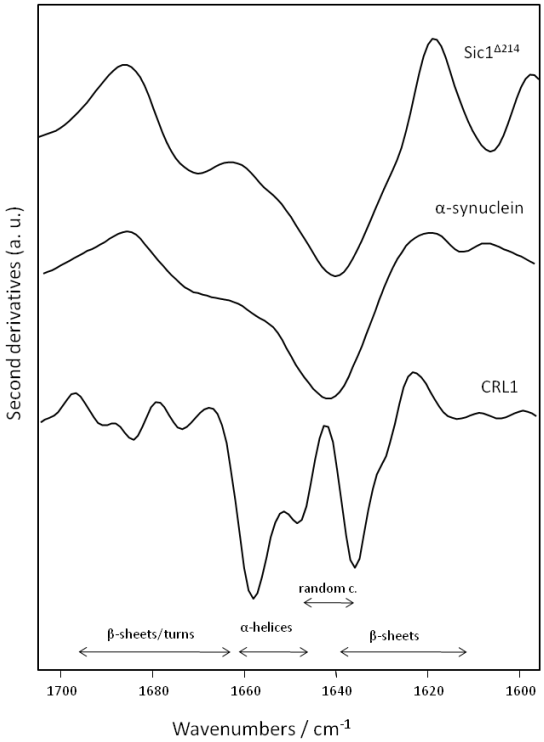


Figure S1. FT-IR analysis of folded and intrinsically disordered proteins. Secondderivative FT-IR spectra of Sic1$^{\Delta 214}$, α-synuclein (13), and Lipase 1 of Candida rugosa (CRL1) (36). The spectrum of Sic1□214 is very similar to that of the IDP α-synuclein. The absorption regions of secondary-structure types are indicated below the spectra. The second derivatives were calculated after normalization of the measured absorption spectra by the Amide-I band area.

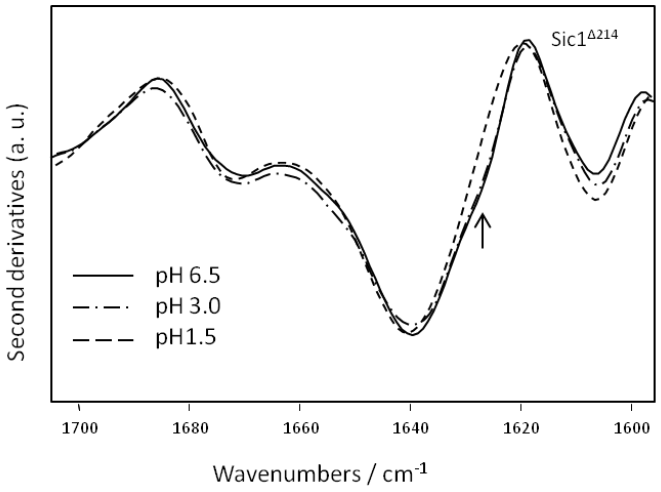


Figure S2. Effect of pH on Sic1$^{\Delta 214}$ secondary structure. FT-IR second-derivative spectra of Sic1$^{\Delta 214}$ at pH 6.5, 3.0 and 1.5 are compared. The arrow points to the decrease of pH.

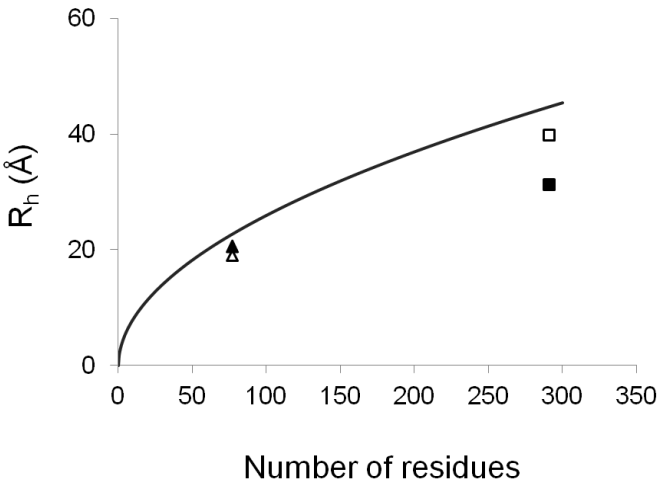


Figure S3. Comparison of experimental and calculated hydrodynamic radii. The curve was built from equation (2) of reference (34). Sequence-specific values for Sic1$^{\Delta 214}$ (triangles) and Sic1^{FL} (squares), calculated by equation (6) of reference (34), are plotted as open symbols (see also Table I). Experimental values from gel-filtration chromatography are plotted as black solid symbols.

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

Frontiers in Physiology. 2012;3:435.

Matteo Lambrughi†, Elena Papaleo*†, Lorenzo Testa, Stefania Brocca, Luca De Gioia and Rita Grandori*

Department of Biotechnology and Biosciences, University of Milano-Bicocca, Milan, Italy

†These authors contributed equally to this work.

* Correspondence: ElenaPapaleo and RitaGrandori, Department of Biotechnology and Biosciences, University of Milano-Bicocca, Piazza della Scienza2, Milan, Italy.
e-mail: elena.papaleo@unimib.it; rita.grandori@unimib.it

ABSTRACT

Cyclin-dependent kinase inhibitors (CKIs) are key regulatory proteins of the eukaryotic cell cycle, which modulate cyclin-dependent kinase (Cdk) activity. 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 (3D) structure in their free state, undergoing folding upon binding to specific partners. Unbound IDPs are not merely random-coil structures, but can present intrinsically folded structural units (IFSUs) and collapsed conformations. These structural features can be relevant to protein function *in vivo*. The yeast CKI Sic1 is a 284-amino acid IDP that binds to Cdk1 in complex with the Clb5,6 cyclins, preventing phosphorylation of G1 substrates and, therefore, entrance to the S phase. Sic1 degradation, triggered by multiple phosphorylation events, promotes cell-cycle progression. Previous experimental studies pointed out a propensity of Sic1 and its isolated domains to populate both extended and compact conformations. The present contribution provides models for compact conformations of the Sic1 kinase-inhibitory domain (KID) by all-atom molecular dynamics (MD) simulations in explicit solvent and in the absence of interactors. The results are

integrated by spectroscopic and spectrometric data. Helical IFSUs are identified, along with networks of intramolecular interactions. The results identify a group of putative hub residues and networks of electrostatic interactions, which are likely to be involved in the stabilization of the globular states.

Keywords: intrinsically disordered proteins, Sic1, electrostatic interactions, cyclin-dependent kinase, molecular dynamics simulations, electrospray ionization mass spectrometry

INTRODUCTION

Intrinsically disordered proteins (IDPs) are a class of promiscuous proteins that do not possess a well-defined three-dimensional (3D) structure in solution. Several IDPs or disordered domains can fold into ordered structures upon interactions with binding partners (Dyson and Wright, 2005; Espinoza-Fonseca, 2009b), although cases are known in which structural disorder is retained also in the bound state (Tompa and Fuxreiter, 2008; Meszaros et al., 2011). IDPs are very common in nature and notably in eukaryotes, where ~50% of the proteins are predicted to contain long disordered regions, and ~30% are classified as IDPs (Oldfield et al., 2005; Uversky and Dunker, 2010). These data are consistent with the observation that IDPs often play key regulatory roles in biological processes (Uversky et al., 2000; Uversky, 2002; Dyson and Wright, 2005; Tompa, 2005; Uversky and Dunker, 2010).
Most IDPs can transiently populate partially structured conformations in their unbound state and display intrinsically folded structural units (IFSUs). These elements are more often helical and are thought to provide seeds for binding interfaces (Sivakolundu et al., 2005; Espinoza-Fonseca et al., 2007, 2012; Belle et al., 2008; Espinoza-Fonseca, 2009a; Wright and Dyson, 2009; Kjaergaard et al., 2010; Norholm et al., 2011). Moreover, unbound IDPs can populate collapsed, globular conformations, stabilized by intramolecular interactions of both electrostatic and hydrophobic nature (Marsh et al., 2007; Espinoza-Fonseca, 2009a; Wostenberg et al., 2011).
Considerable efforts have been devoted to model IDP structural and dynamic properties at the atomic level (Dyson and Wright, 2005; Dunker et al., 2008; Turoverov et al., 2010; Fisher and Stultz, 2011). The understanding of IDP molecular recognition requires 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). Atomistic or coarse-grained molecular dynamics (MD) simulations have proven suitable to describe the conformational landscape of IDPs or denatured proteins, identifying intrinsic or residual structure (Espinoza-Fonseca et al., 2007; Espinoza-Fonseca, 2009a, 2012; Yoon et al., 2009; Rauscher and Pomes, 2010; Cino et al., 2011; Fisher and Stultz, 2011; Qin et al., 2011; Ganguly et al., 2012; Lindorff-Larsen et al., 2012).

Several IDPs are involved in cell-cycle regulation (Galea et al., 2008). Among these, cyclin-dependent kinase inhibitors (CKIs) modulate the activity of cyclin-dependent kinases (Cdks) playing a key role in the regulation of the eukaryotic cell cycle (Besson et al., 2008; Barberis, 2012). 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). The G1-to-S transition is executed only upon ubiquitin-dependent Sic1 degradation by the proteasome. Sic1 is targeted for destruction by the Skp1-Cul1-F-box complex, SCFcdc4. 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). 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).

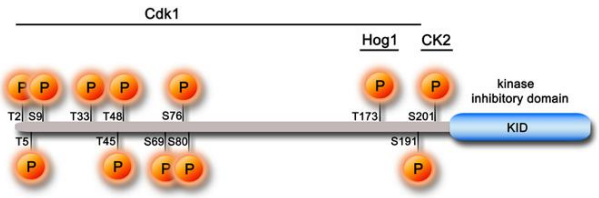


Figure 1. Schematic representation of the full-length Sic1 protein and its principal phosphorylation sites. The minimal functional KID fragment, corresponding to the last 70 residues of the protein, is indicated as a blue box. The main kinases that regulate Sic1 activity in the cell cycle are indicated,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

Sic1 is 284-residue long and, although disordered in its whole length, it can populate 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, 2011b). Its free state is characterized by tertiary and secondary structures (Brocca et al., 2009). The little content in secondary structures (mainly helical) is distributed quite uniformly throughout the polypeptide chain, although the C-terminus is slightly more ordered than the N-terminus (Brocca et al., 2011a,b). 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 (Toyoshima and Hunter, 1994; Barberis et al., 2005) (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).

This work investigates the conformational ensemble of Sic1 KID by integrating all-atom explicit solvent MD simulations with experimental data to achieve a model of the compact conformations populated by the protein in the absence of interactors. The results identify regions that are likely to be characterized by intrinsic secondary and tertiary structure and point out a predominant role of electrostatic interactions in promoting protein compaction.

MATERIALS AND METHODS

Starting Structures for MD Simulations

An extended conformation of the Sic1 KID fragment (residues 215–284) was modeled 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 were generated by the program MODELLER, including constraints derived from the prediction of secondary structure (Barberis et al., 2005). These models feature a main-chain root mean square deviation (RMSD) between 0.7 and 10 nm in pairwise comparisons. 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) were constrained to α-helices (T226-L238 and I244-I248). In fact, MODELLER can be used to generate 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 300 K 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 pair 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 analyzed 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 analyzed using a 0.6 nm distance cutoff.

Analysis of Salt-Bridge Interaction Networks

Salt-bridge interactions were also analyzed by the *Pymol* plugin *xPyder*, representing pairwise relationships extracted from protein structures by 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 in the MD ensemble (i.e., the number of trajectory frames in which the interaction was present divided by the total number of frames) was 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.
Residues connected by more than 3 edges to their neighbors (Brinda and Vishveshwara, 2005) are referred to as hubs of the interaction network. Hubs in protein networks are known to play key roles in protein structure and 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-only matrices were calculated for the macro-trajectory. Given a reaction coordinate qα, the probability of finding the system in a particular state $q\alpha$ is proportional to ($e^{-G(q\alpha)/kT}$), where $G(q\alpha)$ is the Gibbs free energy of that state. The FEL can be estimated 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 root mean square inner product (RMSIP) is a measure of similarity between subspaces sampled by different trajectories and it was calculated for each pair of independent replicates, comparing the subspace described by the first 10 eigenvectors from PCA (Amadei et al., 1999) according to the formula:

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

where $\eta^A{}_i$ and $\eta^B j$ are the eigenvectors to be compared and D the number of eigenvectors considered.

Cluster Analysis

The structures belonging to each FEL basin were isolated and the pairwise main-chain RMSD matrix was calculated for each basin. 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):

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

where $D^2(s_i, s_j)$ is the Cα mean-square distance (MSD) between the structures s_i and s_j, and $\langle D^2 \rangle$ 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 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

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 His6 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 of 1 μm internal diameter. The instrument was calibrated using the 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 V, declustering potential 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).

RESULTS

Structural models of Sic1 KID (residues 215–284) in an extended conformation were generated as described in the "Materials and Methods", satisfying constraints on secondary structure according to the previously published predictions (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 analyzed 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; Zhuravlev et al., 2009).

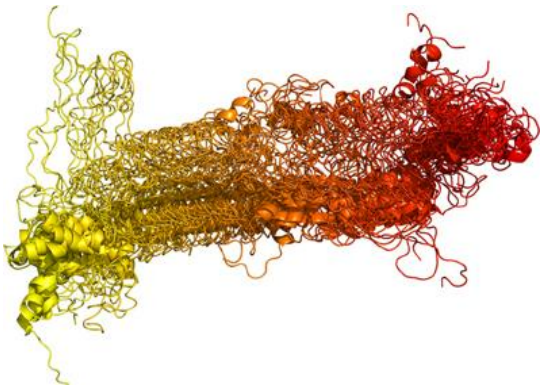


Figure 2. Structural displacement of the Sic1 KID domain in the MD macro-trajectory. The equilibrated parts of each replicate were concatenated in a single macro-trajectory of more than 0.5 μs, shown here. Snapshots are collected every 2 ns, overlaid and represented by a color gradient, from yellow (the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

Order Parameter O

The order parameter O was calculated to evaluate the heterogeneity of the conformational ensemble described by the MD simulations (Fisher and Stultz, 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.

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-only 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 can describe the sampled conformational landscape, if combined to FEL calculations. The projections of the MD macro-trajectory on the first two PCs show an efficient sampling of the conformational space. In fact, distinct simulations sample different regions of the subspace described by the first two PCs, featuring partial overlap (Figure 3A). In particular, simulations starting from different initial models often populate the same basins, confirming that the MD ensemble provides a good description of Sic1 KID conformational 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 PCs, as well as along the second and the third ones (Figures 3B,C).

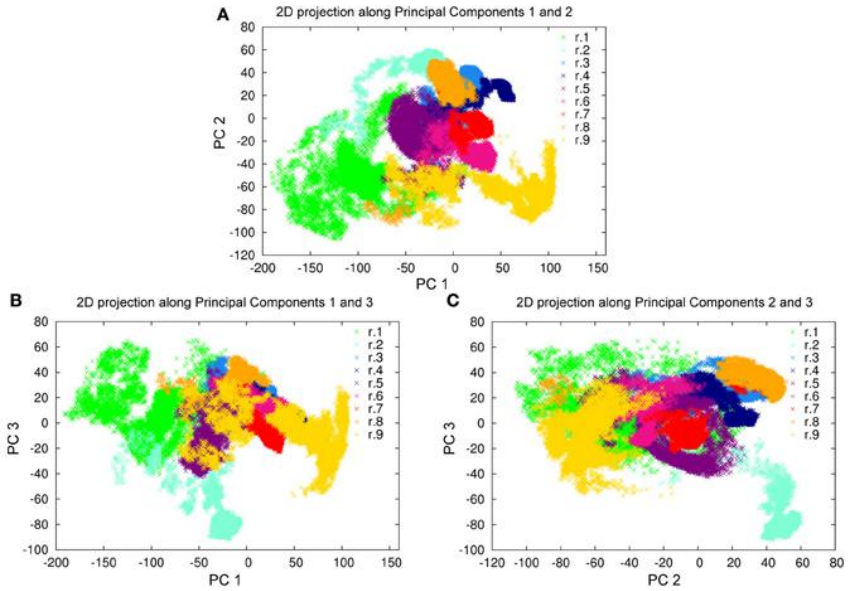


Figure 3. Projections of the MD trajectories of Sic1 KID on the first 3 PCs derived from PCA of the macro-trajectory. (**A**) First (*x* axis) and second (*y* axis) eigenvectors, (**B**) first (*x* axis) and third (*y* axis) eigenvectors, (**C**) second (*x* axis) and third (*y* axis) eigenvectors. The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

To better quantify the trajectory overlap, the RMSIP of the first ten PCs between the different replicates (Figure 4A) was calculated as described in the "Materials and Methods". In fact, RMSIP is an index of similarity between essential subspaces. In our simulations, the independent replicates with highest overlap have a RMSIP value higher than 0.5, and the average RMSIP is 0.44. Overall, the RMSIP analysis indicates satisfactory convergence, with a high overlap of the dynamics information provided by the simulations.

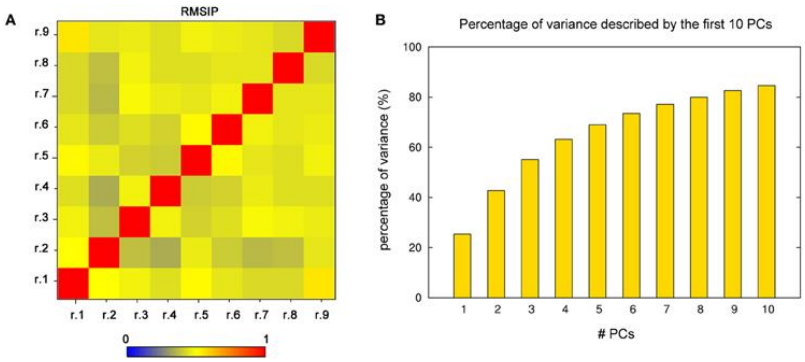


Figure 4. PCA analysis. (**A**) RMSIP matrix calculated for each pairwise comparison of the single replicates, using the conformational space described by the first 10 PCs. The RMSIP values are highlighted by a color gradient, from blue to red. (**B**) Cumulative percentage of the variance described by the first ten PCs. The first three (or five) PCs account together for more than 50% (or 70%) of the global motions of the domain.

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 PCs 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 as reaction coordinates (Figure 5). The average structure representative of each FEL basin is also reported in Figure 5. 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 applied to classical MD simulations are not sufficiently accurate to calculate the free-energy barriers and to describe the location of metastable states and barriers in details (Altis et al., 2008). Nevertheless, they can be 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, in line with the data reported in Figure 3.

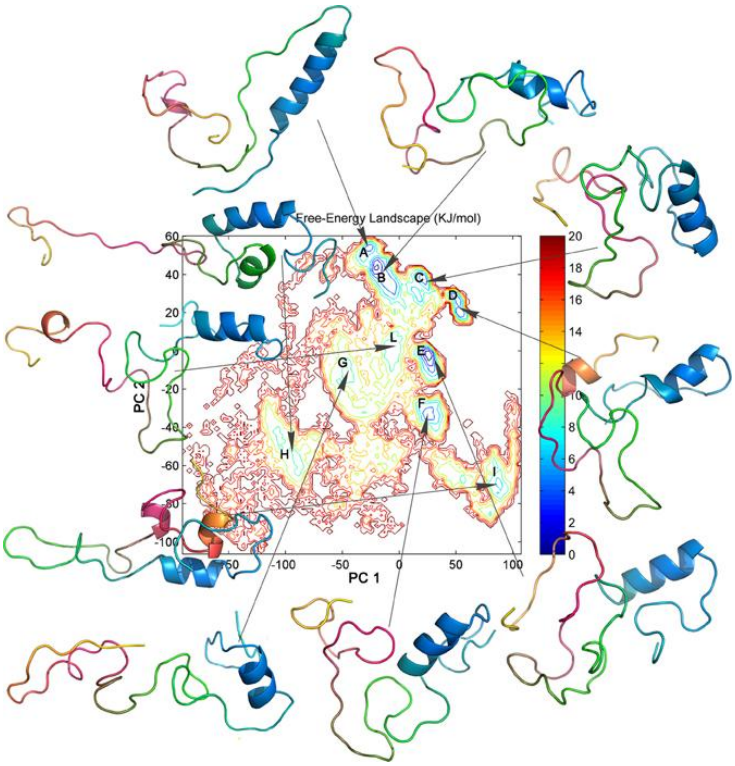


Figure 5. FEL representation calculated using the first two PCs as reaction coordinates. The major basins are labeled by capital letters (A to L). The free energy is given in kJ/mol and indicated by the color code shown on the figure. The average 3D structure identified for each conformational basin by structural clustering is represented as cartoon. The 3D structures are highlighted by a color gradient, from N-terminus (cyan) to C-terminus (yellow).

For the structures of 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, 3). The 3D 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.

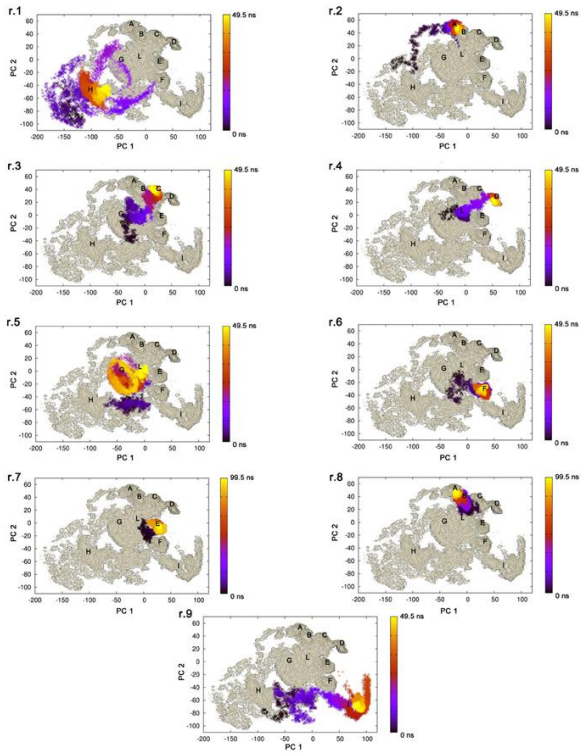


Figure 6. Projection of the single replicate on the FEL. The trajectories explored by each replicate are depicted on the FEL representation obtained using the first two PCs as reaction coordinates. The temporal evolution of each replicate is indicated by a color gradient, from black (0 ns) to yellow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

Table 2. Secondary structure content and SAS values.

Basin name	Helical content average (α–3_{10}–π)	Helical content maximum (α–3_{10}–π)	Total helical content average (%)	SAS average (nm^2)	SAS minimum (nm^2)	SAS maximum (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

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 Sic1KID fragment. The minimum, maximum, and average SAS values are also displayed.

Table 3. Hub residues identified for each conformational basin.

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

The residues involved in three or more connections in the salt-bridge networks have been considered as hubs.

The here described principal motions of Sic1 KID identify distinct N- and C-terminal sub-domains. The first PC mainly accounts for the dynamics of the C-terminal region (yellow and hot-pink in the snapshots shown in Figure 5), whereas the second PC 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 in our MD ensemble. In fact, these motions are related to the pairing of the C-terminal and N-terminal regions and are likely to reflect the transitions between an open and a closed state of the domain in the MD framework described here. The closed state is characterized by a long unstructured loop in the central region of the domain, approximately spanning residues 243–270 (green in the snapshots shown in Figure 5).

Secondary Structure Content

The MD data were analyzed 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, 310, 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 structures of each 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, 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). 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 upon acidification, 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 common to the current force fields (Click et al., 2010; Knott and Best, 2012).

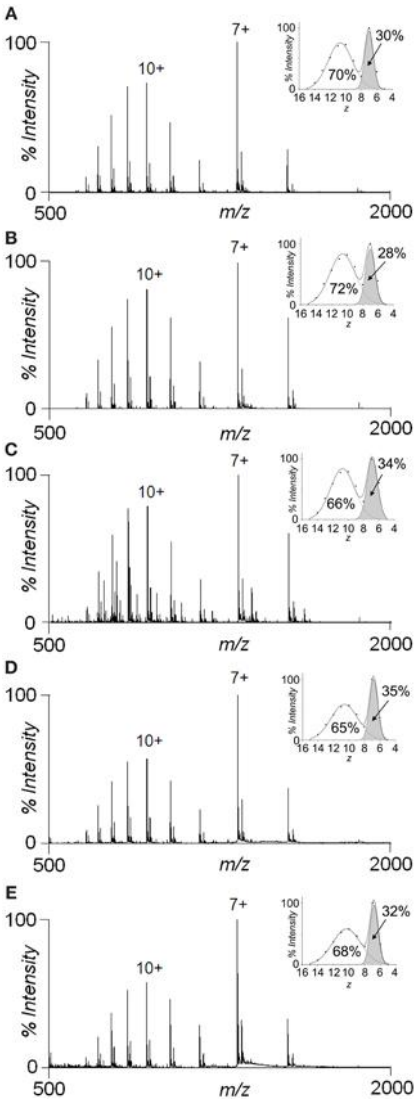


Figure 7. Species distributions by ESI-MS and effects of organic solvents. Nano-ESI-MS spectra of 10 μM protein in **(A)** 50 mM ammonium acetate, pH 6.5; **(B)** 50 mM ammonium acetate, pH 6.5, 30% acetonitrile; **(C)** 50 mM ammonium acetate, pH 6.5, 50% acetonitrile; **(D)** 50 mM ammonium acetate, pH 6.5, 30% methanol; and **(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.

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). 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 nm2 and corresponds to the compact conformations discussed above (from basins B, C, D, E, F).

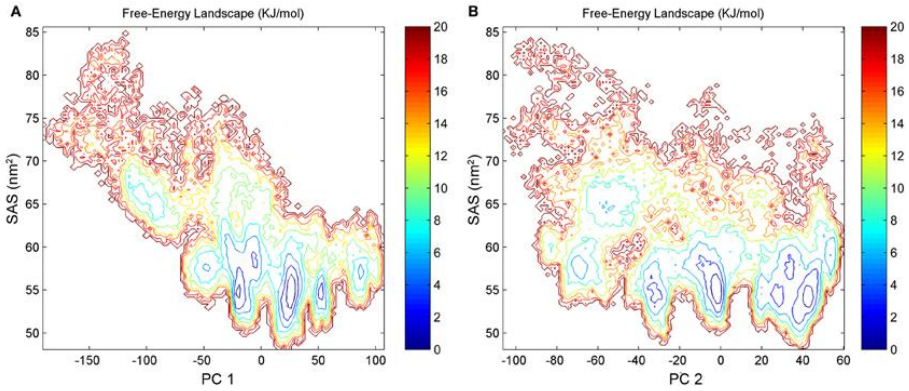


Figure 8. FEL representations calculated using the first **(A)** or the second **(B)** PCs and SAS values as reaction coordinates. The free energy is given in kJ/mol and indicated by the color code shown on the figure.

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 analyzed 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. In spite of the presence of several charged residues, often consecutive in the amino acid sequence, it also displays quite low net charge per residue (Brocca et al., 2011b). These features are consistent with the propensity displayed by Sic1 KID for globular states (Mao et al., 2010; Brocca et al., 2011a,b).

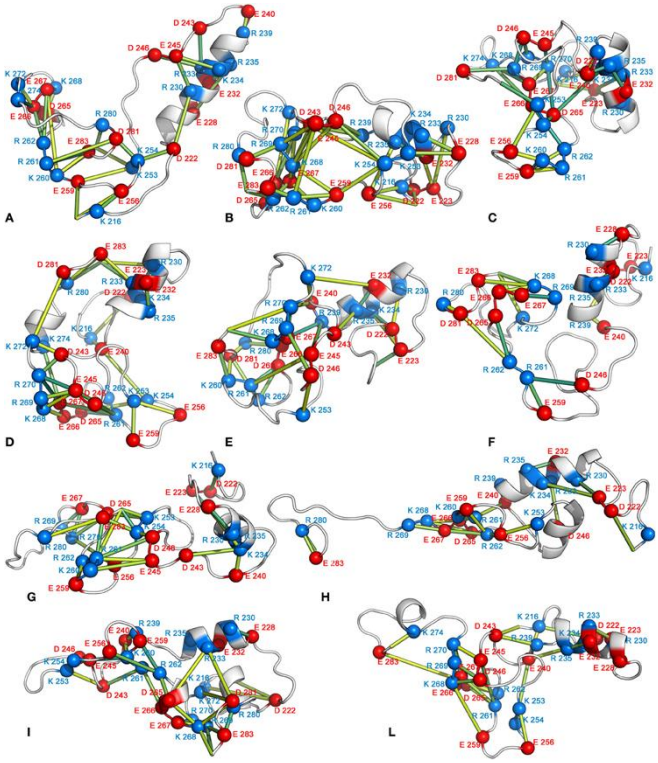


Figure 9. Networks of salt bridges. The 3D average structure of each FEL basin is shown as cartoon, and the Cα atoms of the acidic and basic residues involved in salt bridges are shown as red and blue spheres, respectively. The Cα atoms of the interacting residues are connected by sticks of different shades of color depending on the interaction persistence (from yellow to green for increasing persistence values).

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, 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 (data not shown). 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).

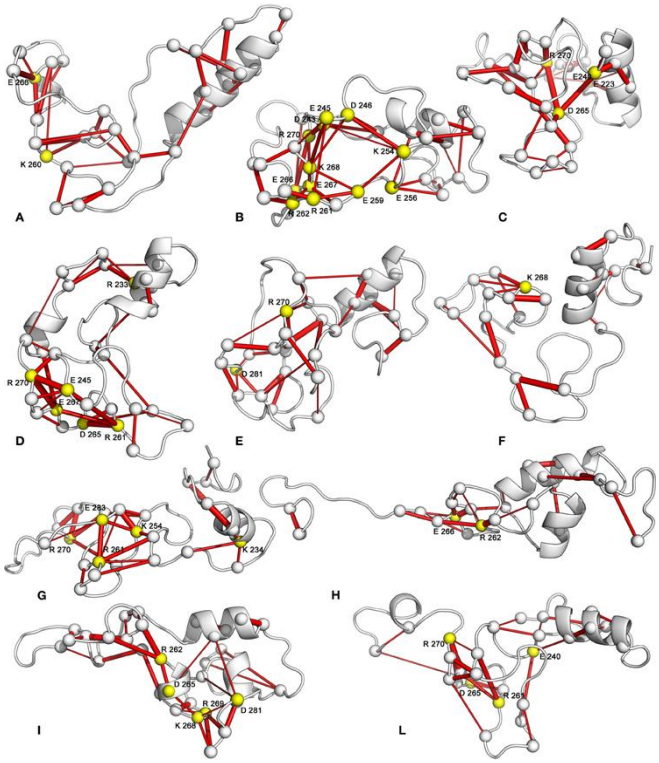


Figure 10. Hub residues in salt-bridge networks. The 3D average structure of each FEL basin is shown as cartoon, and the Cα atoms of the residues involved in salt bridges are shown as spheres. The Cα atoms of the interacting residues are connected by sticks, whose thickness is proportional to the persistence of the interaction. Hub residues, defined as those involved in at least three different salt-bridge interactions, are highlighted in yellow.

Sub-networks of salt bridges have also been identified (Figure 11). Less compact states show a greater number of small 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.

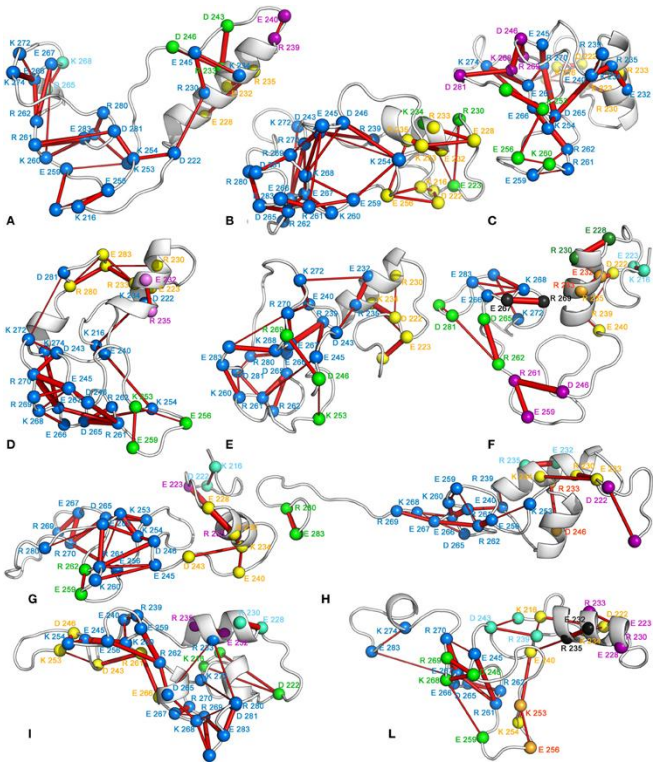


Figure 11. Sub-networks of salt bridges. The 3D average structure of each FEL basin is shown as cartoon, and the Cα atoms of the residues involved in salt bridges are shown as spheres. The Cα atoms of the interacting residues are connected by sticks, whose thickness is proportional to the persistence 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 the smallest one (2 residues).

Some amino-aromatic interactions could also be detected, even if generally characterized by very low persistence. 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).

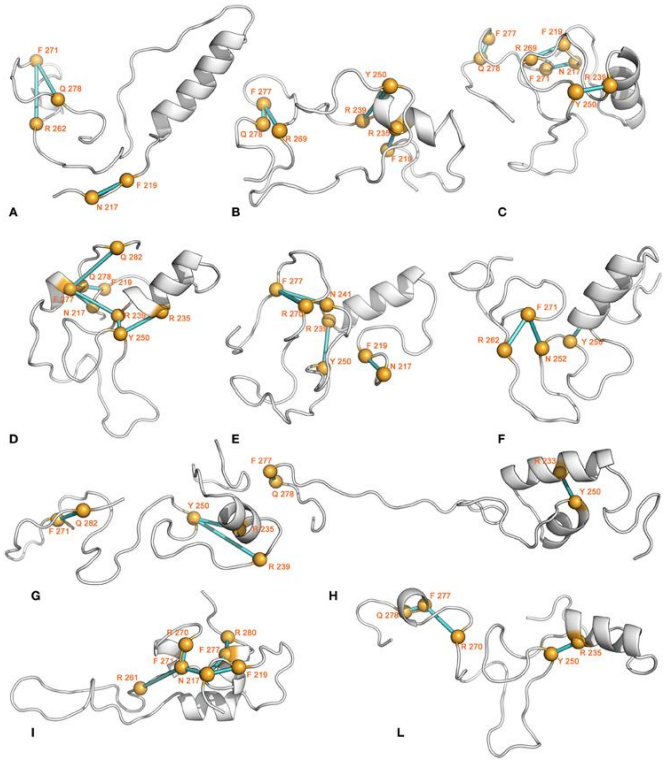


Figure 12. Amino-aromatic interactions. The 3D average structure of each conformational basin is shown as cartoon, and the Cα atoms of the residues involved in amino-aromatic interactions are shown as yellow spheres. The Cα atoms of the interacting residues are connected by sticks of different shades of color, depending on the interaction persistence (from cyan to blue for increasing persistence).

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).

DISCUSSION

The heterogeneous and dynamic nature of IDPs makes structural characterization of their unbound state highly challenging. Although MD force fields have been developed to simulate protein folding, they have also proven 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 computational results are complemented by biophysical data. These studies contribute to enforce the applicability of classical MD to complex molecular ensembles. Nevertheless, analysis 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 employ here 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 dynamic.

The present results could also be interpreted in the light of the structural and functional relation to the mammalian p21 and p27 KID domains (Barberis et al., 2005). In fact, it has been shown that p27 can replace Sic1 in yeast cells (Barberis et al., 2005) and that Sic1 KID can functionally interact with mammalian Cyclin A-Cdk2 inhibiting 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 point to 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 α-helical 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 of binding (Otieno et al., 2011). The corresponding and structurally similar region of Sic1 KID might, therefore, promote binding to the cyclin/kinase 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.

Moreover, according to the MD scenario reported here, electrostatic interactions seem to be the major determinant of structural compaction in the isolated Sic1 KID. This result of the MD simulations is supported by experimental evidence by ESI-MS. 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).

Electrostatic interactions could also be 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.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ACKNOWLEDGMENTS

This work was supported by the Standard HPC Grant 2011 and 2012 from CASPUR to Elena Papaleo, by a grant FAR (Fondo Ateneo per la Ricerca) of the Università Milano-Bicocca to Rita Grandori, Luca De Gioia, Stefania Brocca and by a doctoral fellowship "Fondazione Fratelli Confalonieri" to Lorenzo Testa. 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.

REFERENCES

Altis A., Otten M., Phuong H., Nguyen P. H., Hegger R., Stock G. (2008). Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis. J. Chem. Phys. 128, 245102.

Amadei A., Ceruso M. A., 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., Berendsen H. J. (1993). Essential dynamics of proteins. Proteins 17, 412–425.

Angelova K., Felline A., Lee M., Patel M., Puett D., 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., 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. 279, 3386–3410.

Barberis M., De Gioia L., Ruzzene M., Sarno S., Coccetti P., Fantucci P., et al. (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., 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., et al. (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., et al. (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., Svergun D. I. (2012). Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. Mol. Biosyst. 8, 151–167.

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

Brinda K. V., 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., et al. (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., 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., et al. (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. Nat. Protoc. 2, 2728–2733.

Cino E. A., Wong-ekkabut J., Karttunen M., 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., 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., et al. (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., Clifford S. (2009). Introduction to Algorithms. 3rd Edn Vol. 1, Cambridge, MA: MIT Press, 1292.

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, 10089–10092.

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

Dobo A., 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., Sussman J. L. (2008). Function and structure of inherently disordered proteins. Curr. Opin. Struct. Biol. 18, 756–764.

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

Escote X., Zapater M., Clotet J., 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., Ilizaliturri-Flores I., Correa-Basurto J. (2012). Backbone conformational preferences of an intrinsically disordered protein in solution. Mol. Biosyst. 8, 1798–1805.

Espinoza-Fonseca L. M., Kast D., 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., 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.

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

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

Fisher C. K., Stultz C. M. (2011). Constructing ensembles for intrinsically disordered proteins. Curr. Opin. Struct. Biol. 21, 426–431.

Fuhrmans M., Sanders B. P., Marrink S. J., 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., 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., Chen J. (2012). Synergistic folding of two intrinsically disordered proteins: searching for conformational selection. Mol. Biosyst. 8, 198–209.

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

Hazy E., 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., Fraaije J. (1997). LINCS: a linear constraint solver for molecular simulations. J. Comput. Chem. 18, 1463–1472.

Hodge A., 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., Lewis R. J. (2001). Structural basis for control by phosphorylation. Chem. Rev. 101, 2209–2242. doi: 10.1021/cr000225s.

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

Kaltashov I. A., 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., van Gunsteren W. F. (2010). Comparing geometric and kinetic cluster algorithms for molecular simulation data. J. Chem. Phys. 132, 074110.

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, 12535–12540.

Knott M., 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., et al. (2011). Cascades of multisite phosphorylation control Sic1 destruction at the onset of S phase. Nature 480, 128–131.

Lindorff-Larsen K., Trbovic N., Maragakis P., Piana S., 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., 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., 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., et al. (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., 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., 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., Dosztanyi Z. (2011). The expanding view of protein-protein interactions: complexes involving intrinsically disordered proteins. Phys. Biol. 8, 035003.

Mittag T., Orlicky S., Choy W. Y., Tang X., Lin H., Sicheri F., et al. (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., et al. (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., et al. (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., 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., Dunker A. K. (2005). Comparing and combining predictors of mostly disordered proteins. Biochemistry 44, 1989–2000.

Otieno S., Grace C. R., 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., 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., Papaleo E. (2012). xPyder: a PyMOL plugin to analyze coupled residues and their networks in protein structures. J. Chem. Inf. Model. 52, 1865–1874.

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

Rauscher S., 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., Longhi S. (2006). Assessing protein disorder and induced folding. Proteins 62, 24–45.

Russo A. A., Jeffrey P. D., Patten A. K., Massague J., 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., et al. (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., et al. (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., 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., 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., Tompa P. (2011). Protein disorder prevails under crowded conditions. Biochemistry 50, 5834–5844.

Testa L., Brocca S., 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., Grandori R. (2011b). Electrospray ionization-mass spectrometry conformational analysis of isolated domains of an intrinsically disordered protein. Biotechnol. J. 6, 96–100.

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

Toyoshima H., 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., 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. doi: 10.1110/ps.4210102.

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

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

Verma R., Feldman R. M., 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., Hansia P. (2009). Intra and inter-molecular communications through protein structure network. Curr. Protein Pept. Sci. 10, 146–160.

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

Wright P. E., 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., et al. (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., 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., Papoian G. A. (2009). Deconstructing the native state: energy landscapes, function, and dynamics of globular proteins. J. Phys. Chem. B 113, 8800–8812.

Heterologous expression, purification and characterization of Sic1-physiological interactors, Cdk1 and Clb5

Preliminary results

Lorenzo Testa, Rita Grandori, Stefania Brocca

ABSTRACT

The G1-to-S transition during the mitotic cell cycle in the yeast *Saccharomyces cerevisiae* requires the activity of the Cdk1-Clb5/6 complexes. These complexes are inhibited by the intrinsically disordered protein (IDP) Sic1 till the end of the G1 phase, when the inhibitor is degraded and cells can be driven into S phase. This work reports expression and purification of Cdk1 and Clb5. Cdk1 is expressed in a soluble form in *E. coli* cells, and hence can ben purified in the proper quality and amount to perform a preliminary structural description. Full-length Clb5, instead, is not expressed efficiently, probably because of an N-terminal disordered region that favors proteolytic events. A truncated form of the protein, lacking the first 156 amino acids ($Clb5^{\Delta 156}$) turns out to be expressed, although it displays a high propensity to aggregate into inclusion bodies (IBs). A procedure for solubilization of IBs and on-column refolding of $Clb5^{\Delta 156}$ is developed to produce the cyclin fragment in a soluble form.

Introduction

The transition from the G1-to-S phase during the mitotic cell cycle in the yeast *Saccharomyces cerevisiae* requires the activity of the Cdk1-Clb5/6 complexes [1]. Sic1 is a kinase inhibitory protein (KIP) playing a key role in this transition, where it acts as a cyclin-dependent protein kinases inhibitor (CKI) by forming ternary complexes with kinases and their cognate cyclins. In particular, Sic1 controls the timing of entrance in S phase during the yeast cell cycle by inhibiting the Cdk1-Clb5/6 complexes, whose activity is required for the G1-to-S transition [2]. At the end of the G_1 phase Sic1 is degraded allowing Clb5,6/Cdc28 to drive cells into S phase [3]. It is now known that Sic1 must be phosphorylated at multiple site in its N-terminal region by the Cln1,2/Cdc28 complexes before it can be ubiquitinated and degraded [4-5]. Sic1 is a 284-amino acid protein, previously identified as an IDP [6]. The last 70 residues, being the minimal protein fragment for *in vivo* Cdk1 inhibition, have been identified as the kinase inhibitor domain (KID) [7]. The conformational ensemble of Sic1 KID has been investigated by integrating all-atom explicit solvent molecular dynamics (MD) simulations with experimental data [8-9], in order to achieve a model of the compact conformations populated by the protein in the absence of interactors. This study identified the regions likely to be characterized by intrinsic secondary and tertiary structure and pointed out a predominant role of electrostatic interactions in promoting protein compaction [9].

Despite the molecular details concerning Sic1 destruction, in particular the mechanisms involving Sic1 N-terminal region and its ubiquitination, have been elucidated [10-11], no structural data are available for the Sic1-Cdk1-Clb5/6 ternary complex, so the molecular mechanism by which Sic1 inhibits S-Cdk1 activity remains obscure. On the contrary, an in-depth functional and structural characterization of p21 and p27/KIP1, which are IDPs involved in the regulation of mammalian cell cycle, have been achieved [12]. Sic1 was initially proposed as a functional homologue of p21 [13], which in turn is characterized by significant sequence similarity with p27 (42% identity) [14]. The inhibitory mechanism of p27, investigated by several biophysical techniques [15-16], is a stepwise process which has different consequences: p27 occupies a secondary substrate recruitment site on cyclin A with one of its subdomains [17]; it binds to the N-terminal lobe of Cdk2 with a second subdomain, flattening the Cdk2-lobe out and disrupting the active site; it blocks ATP binding to Cdk2 [18]. It has been suggested that the inhibitor acts as a structural and thermodynamic staple in the ternary complex [19]. It can be hypothesized that Sic1 interacts with the Cdk–cyclin complex by a

similar two-step mechanism [14]. To test such an hypothesis, it would be very useful to identify the inter-residues contacts responsible for the interaction, in order to compare them with the mammalian ternary complex and to disclose the evolutionary relations between yeast and mammalian inhibitors. Moreover, the thermodynamic description of the Sic1-Cdk1-Clb5/6 complex, in terms of dissociation constant, energy contributions, and binding kinetics (association and dissociation rates), would contribute to the cell-cycle modeling based on the systems-biology approach [20-21]. Finally, since mammalian regulator p21 and p27 are regarded as desirable drug targets against cancer [22], a better investigation of their homologue in the model system *S. cerevisiae* could provide intriguing perspectives for the development of future therapies.
In the present work we report our results on the heterologous expression and purification of Cdk1 and Clb5, as a preliminary step of a structural study on the complex these proteins form with their physiological interactor Sic1.

Materials and Methods

Sequence analyses

The primary sequences of Cdk1 and Clb5$^{\Delta 156}$ were investigated to identify aggregation-prone regions by different methods: Tango [23], Aggrescan [24] and Waltz [25]. The analyses were carried out using the default setting of the parameters for each algorithm.
Disorder predictions for the full-length Clb5 were performed by different predictors of naturally disordered regions (PONDRs), a series of feedforward neural networks that use sequence attributes taken over windows of 9 to 21 amino acids. These attributes, such as the fractional composition of particular amino acids, hydropathy, or sequence complexity, are averaged over these windows and the values are used to train the neural network during predictor construction. The same values are used as inputs to make predictions [26]. In particular, the Clb5 sequence was analyzed by PONDR VL3 [27], PONDR VL-XT [28], PONDR XL1-XT [27] and PONDR CaN-XT [29].

Strains and Growth Media

Escherichia coli strain DH5α (Invitrogen, Life Technologies Corporation, Carlsbad, California) was used as the host for DNA amplification, whereas strain BL21-Rosetta 2(DE3) (Novagen, EMD Chemicals Inc, Darmstadt,

DE) was used for heterologous expression. *E. coli* cells were grown in low-salt Luria–Bertani (ls-LB) medium (10 g peptone, 5 g yeast extract, and 5 g NaCl in 1 l water). *E. coli* transformants were selected on ls-LB agar plates supplemented with the appropriate antibiotics [34 mg/l chloramphenicol and 50 mg/l ampicillin for BL21-Rosetta 2(DE3), and 50 mg/l ampicillin for DH5α].

Cloning

Standard recombinant-DNA techniques were applied according to Sambrook et al. [30]. Restriction enzymes and DNA ligase were from New England Biolabs (NEB, Ipswich, MA, USA), while the primers were purchased from Sigma-Aldrich (St. Louis, MO, USA). The CLB5 encoding sequence from pET21[CLB5]$_{His6}$ was subjected to mutagenic PCR to obtain deletions of its 5'-moiety. Deletions were carried out by back-to-back PCR [31], with non-overlapping phosphorylated oligonucleotides. The primers were designed to amplify the whole plasmid containing the target gene, excluding the DNA fragment to be deleted. After PCR, the blunt-ended amplimers were directly ligated resulting in circularization of the plasmidic DNA. Primers phosphorylation was carried out before PCR, by incubating oligonucleotides with polynucleotide kinase A (NEB, Ipswich, MA, USA) at 37°C for 1 h. Mutagenic PCRs were carried out in a volume of 50 μl, with 10 ng of pET21[CLB5]$_{His6}$ as a template, with 0.5 μM of each primer. The sequences of the oligonucleotides employed for obtaining the plasmid pET21[CLB5$^{\Delta 156}$]$_{His6}$ were as follows:

- CLB5 $^{\Delta 156}$ forward:
  5'- ATGGCAATGGTAGCAGAATATTCTGCTG - 3'
- CLB5 $^{\Delta 156}$ reverse:
  5'- ATGTATATCTCCTTCTTAAAGTTAAACAAAATTATTTCTAG -3'

PCR amplification was carried out with the high-fidelity PfuII Ultra DNA polymerase (Stratagene, La Jolla, CA) according to the manufacturer's instructions, and applying the following temperature program: 2-min denaturation at 95°C, addition of 1 μl DNA polymerase, 30 cycles of 20 s at 95°C/20 s at 53°C/2 min at 72°C, and a final extension step of 3 min at 72°C. The template DNA was digested by *Dpn*I after amplification. The amplified DNA was purified by ethanol precipitation before self-ligation by T4 DNA ligase. Ligations were carried out with 50 ng DNA in 20 μl reaction mixture at 14°C. The recombinant plasmids were introduced into DH5α or BL21-Rosetta 2(DE3) *E. coli* cells, according to a procedure

adapted from [32], as described by [33]. The final constructs were checked by DNA sequencing (Primm, Milano, Italy).

Protein Expression and Purification

Escherichia coli pre-cultures were inoculated from a single-cell colony from a fresh selection plate. Preparative cultures were obtained in shaking flasks, by 1:20 dilution of pre-cultures with optical density at 600 nm (OD_{600}) between 0.5 and 0.7 in fresh 50 mL autoinduction ZYM-5052 medium [34] and grown at 30°C and 220 rpm for 14-16 hours. The final OD_{600} reached by cultures were ~ 9 for cells carrying the plasmid pET21[CDK1]$_{His}$ and ~ 7 for cells carrying the plasmid pET21[CLB5$^{\Delta 156}$]$_{His}$.
Induced cells were harvested by centrifugation at 9,600 ·*g* at 4°C for 15 min and either immediately used for protein extraction or frozen at -20°C. For the purification of Cdk1, the cell pellet was resuspended in CelLytic B (Sigma-Aldrich St. Louis, MO, USA), a highly efficient yet gentle reagent containing zwitterionic detergents in 40 mM Tris-HCl, pH 8.0. According to the manufacturer's instructions, the ratio of CelLytic B to cell mass was 10-20 mL per gram of wet cell paste/pellet. Cells were lysed by incubating the cell suspension at 4°C for 20 min. The crude extracts were then centrifuged for 10 min at 10,600 ·*g* to collect cellular debris and the insoluble protein fraction. The supernatant was loaded on an immobilized-metal affinity chromatography (IMAC) on Ni^{2+}/NTA beads (Sigma). The column was eluted by 50 mM sodium phosphate, 300 mM NaCl, 250 mM imidazole, pH 8.0. The purified proteins were exchanged to the desired buffer (50 mM sodium phosphate pH 6.5, 300 mM NaCl) by a gel filtration on PD-10 columns (General Electric Healthcare, Little Chalfont, UK). Protein concentration was determined by the Bradford protein assay (Bio-Rad Laboratories, Milano, Italy), using bovine serum albumin as a standard.
Clb5$^{\Delta 156}$ was purified from inclusion bodies (IBs). IBs purification was performed employing the solution of CelLytic B, according to the manufacturer's instructions. The cell pellet was resuspended in CelLytic B, and cells were lysed by incubating the cell suspension at 4°C for 20 min, as for Cdk1 purification. The crude extracts were then centrifuged for 10 min at 10,600 ·*g* to collect cellular debris and the insoluble protein fraction comprising IBs of Clb5$^{\Delta 156}$. The pellet was resuspended in the same volume of CelLytic B used to lyse the cells, and the solution was incubated at room temperature for 10 min. A 10-fold diluted CelLytic B was then added to the suspension, in a ratio of 30 mL per gram of wet cell paste, and the solution was centrifuged at full speed for 5 min at 4°C. The pellet was resuspended

again in a volume of 10-fold diluted CelLytic B equal to that used in the previous step, and the suspension was centrifuged at 10,600 $\cdot g$ for 5 min at 4°C to pellet the IBs. The washing step with 10-fold diluted CelLytic B was repeated twice, to completely remove any remaining soluble proteins and cell wall residues from the IBs.

Isolated IBs were dissolved in 50 mM sodium phosphate, 50 mM Tris pH 12, 8 M urea, at a final concentration of 2 mg/mL, by incubating the suspension at room temperature overnight. Dissolved inclusion bodies were centrifuged for 10 min at 10,600 $\cdot g$ and the pH of the supernatant was lowered to a value of 8.0 with HCl, to allow the binding of proteins on Ni^{2+}/NTA beads manually packed in a mini-column for IMAC purification. The resin was pre-equilibrated with a sodium phosphate buffer at pH 8.0. On-column refolding of denatured $Clb5^{\Delta 156}$ was achieved by washing the column with three volumes of 50 mM sodium phosphate pH 8.0, 300 mM NaCl, 10 mM imidazole and three volumes of 50 mM sodium phosphate pH 8.0, 300 mM NaCl, 20 mM imidazole. The protein was eluted from Ni^{2+}/NTA resin with 50 mM sodium phosphate pH 8.0, 300 mM NaCl, 250 mM imidazole.

SDS-PAGE

SDS-PAGE analyses were carried out on 16% acrylamide Laemmli gels [35]. After electrophoresis, gels were stained by GelCode Blue (Pierce, Illinois, IL, USA). Broad-range, prestained molecular-weight markers (NEB, Ipswich, MA, USA) were used as standards.

Circular dichroism (CD) spectroscopy

The CD spectra were recorded on a spectropolarimeter J-815 (JASCO corporation, Japan), in a 1-mm pathlength cuvette, at 25°C. Samples were in 50 mM sodium phosphate buffer, pH 6.5, 300 mM NaCl. The spectra were acquired with data pitch 0.2 nm, averaged over three acquisitions, andsmoothed by the Savitzky-Golay algorithm [36].

Results and Discussion

Expression and purification of Cdk1

The recombinant protein Cdk1 was expressed in BL21-Rosetta *E. coli* cells transformed with the plasmid pET21[CDK1]$_{HIS6}$ (Codazzi. Unpublished results). Cells were grown in an autoinduction medium containing glucose and lactose, for 16-18 hours at 30°C, cultures at extremely high densities (OD_{600} ~ 9) were obtained. Analyses of expression revealed that under these conditions, a protein with a molecular weight of about 35 kDa was strongly induced (Fig. 1A), as expected taking into account that the theoretical molecular weight of Cdk1-His6 is 35339.86 Da. The overexpressed protein is distributed both in the soluble and insoluble fraction of cellular proteins (Fig. 1A), with the soluble recombinant protein expressed at ~ 30 mg/L.
The amino acid sequence of Cdk1 was scanned by four different algorithms searching for aggregation-prone regions (APRs) likely responsible for its propensity to form inclusion bodies. The methods employed are based on different principles and have been effectively applied to the analysis of aggregation propensity in whole proteomes [37]. Although based on different approaches, all algorithms applied take into account only the primary structure information, considered just a sequential succession of amino acids, and ignore any kind of information on the structure adopted in the non-aggregated/native state. The results obtained, summarized in Fig 1B, are consistent with a low hydrophilic nature of the sequence, providing an explanation for the incomplete solubility of the protein, when overexpressed in *E. coli*. For protein purification, we exploit a gentle lysis buffer containing zwitterionic agents. This procedure allowed us to bring a high fraction of the protein in soluble form, subsequently purified by IMAC. The results of this purification procedure are shown in Fig. 1A, indicating a high purity degree of the final samples. The purified protein was finally subjected to buffer exchange by gel filtration according to the requirements of subsequent analyses. Indeed, biophysical techniques aimed at structural characterization, such as CD and electrospray-ionization mass spectroscopy (ESI–MS), requires low ionic strength, although with different levels of salt tolerance. However, during buffer exchange we observed a loss of protein probably due to the previously identified problems of solubility. The maximal yield was obtained by using a high-salt concentration buffer containing 50 mM sodium phosphate pH 6.5, 300 mM NaCl.

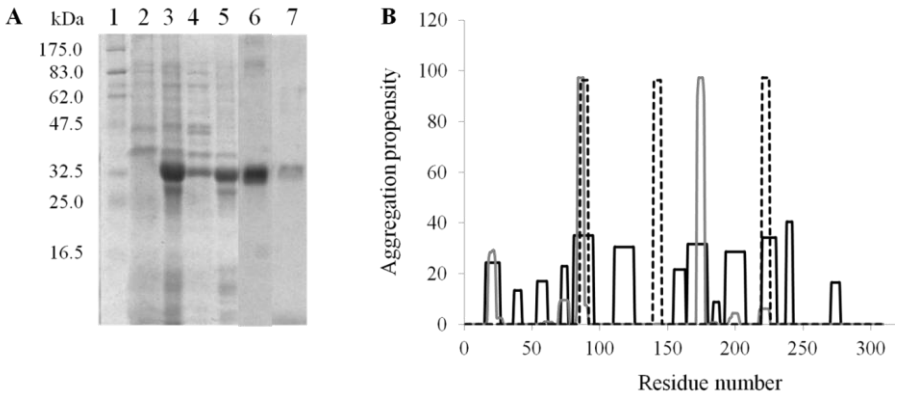


Figure 1. A) Purification of Cdk1 analyzed by Coomassie-stained SDS-PAGE gels. Lane 1: molecular-weight markers; lane 2: total protein content, not induced cells; lane 3: total protein content, induced cells; lane 4: soluble crude extract (sample loaded on IMAC column); lane 5: insoluble crude exctract; lane 6: IMAC-purified protein; lane 7: fraction eluted from gel filtration. (B) Predicted aggregation propensity for Cdk1, as calculated by Aggrescan (black line), Tango (grey line) and Waltz (dotted black line). Since the raw scores obtained with the three softwares are not directly comparable, data were normalized to 100 as the highest score.

Structural Characterization of Cdk1 by CD spectroscopy

Purified Cdk1 was analyzed by CD spectroscopy in the far UV, in order to probe its secondary-structure properties. The spectrum of the protein is reported in Fig. 2. Because of the buffer required to maintain soluble Cdk1, it was not possible to acquire the spectrum below 200 nm. The mean-residue ellipticity is characterized by a negative minimum around 208 nm and a negative shoulder at 222 nm, clearly indicating that the protein contains α-helices. Moreover, the spectrum is quite similar to that of mammalian cyclin-dependent kinase 2 (Cdk2) [38].

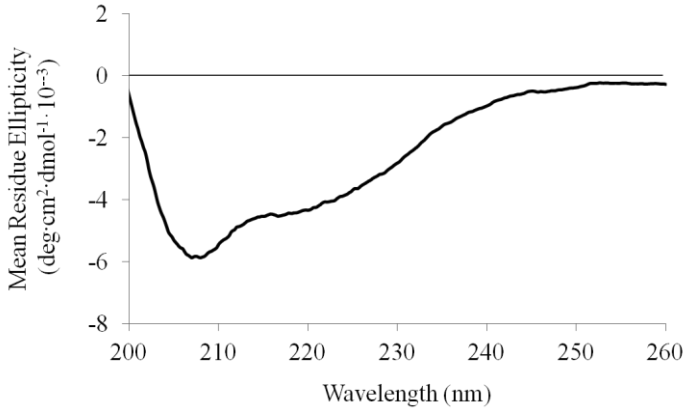


Figure 2. CD spectra in the far UV of 5 μM Cdk1 in 50 mM sodium phosphate pH 6.5, 300 mM NaCl.

Expression and purification of Clb5

Analogously to Cdk1, the expression of the recombinant Clb5 was carried out in BL21-Rosetta *E. coli* cells transformed with the pET21[CLB5]$_{HIS6}$ plasmid (Codazzi. Unpublished results). Unfortunately, no expression was observed, neither after IPTG induction, nor when cells were grown in autoinduction medium (data not shown). The lack of expression has been hypothesized due to an intrinsic instability of the protein. Indeed, the analysis of amino acid sequence performed with different disorder predictors , indicated a highly disordered region in the first 150 amino acids of N-terminal moiety of Clb5 (Fig. 3). Since this region not endowed with a stable tertiary structure is likely to be a target of proteolytic attack [39], we reasoned that the deletion of the N-terminal moiety could allow the expression of Clb5 in a more stable, functional form able to interact with the kinase partner, as it has been already demonstrated for other cyclins expressed in a (correspondingly) truncated form [18, 40].

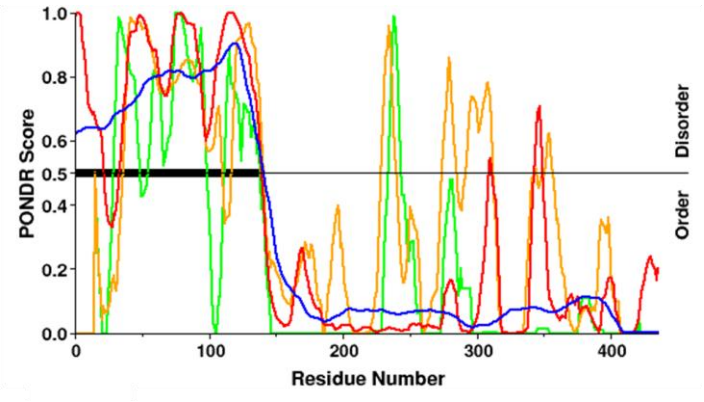


Figure 3. Clb5 disorder analysis. The plot represents the prediction results obtained with PONDR VL3 (blue line); PONDR VL-XT (red line); PONDR XL1-XT (yellow line), PONDR CaN-XT (green line).

On the basis of the disorder prediction and the sequence alignment between Clb5 and human cyclin A (data not shown), we produced a truncated form of Clb5, lacking the first 156 amino acids and therefore named Clb5$^{\Delta 156}$. Unlike the full-length protein, the deleted variant can be expressed by transformed BL21-Rosetta *E. coli* cells either induced with IPTG and grown in autoinduction medium, as shown in Fig 4A. Indeed, analyses of expression clearly indicated an induced protein, with molecular weight comprised between 25 and 32.5 kDa, in good agreement with the theoretical molecular weight of Clb5$^{\Delta 156}$-His6 (33982.21 Da).

When cells carrying the pET21[CLB5$^{\Delta 156}$]$_{HIS6}$ plasmid were grown in the autoinduction medium at 30°C, cultures with optical densities of no more than OD_{600} ~ 7 were obtained. Further investigations about the solubility properties of the overexpressed protein, showed that Clb5$^{\Delta 156}$ is expressed as insoluble protein, at quite high levels (~ 30/40 mg/L) (Fig. 4A). The sequence analysis for the search of APRs, shows a much higher aggregation propensity than Cdk1 (Fig. 4B). The high-level expression of Clb5$^{\Delta 156}$ likely gives rise to inclusion bodies (IBs), resulting from the *in-vivo* accumulation of insoluble aggregates [41]. IBs were isolated by iterative washing steps according to the procedure described in the Materials and Methods section.

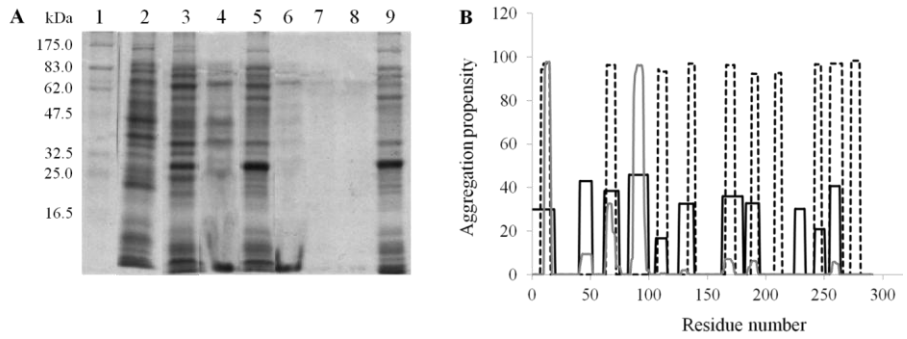


Figure 4. Clb5 aggregation propensity. (A) Purification of $Clb5^{\Delta156}$ analyzed by Coomassie-stained SDS-PAGE gels. Lane 1: molecular-weight markers; lane 2: total protein content, not induced cells; lane 3: total protein content, induced cells; lane 4: soluble crude extract; lane 5: insoluble crude exctract; lane 6: IBs washing step 1; lane 7: IBs washing step 2; lane 8: IBs washing step 3; lane 9: isolated IBs. (B) Predicted aggregation propensity for $Clb5^{\Delta156}$, as calculated by Aggrescan (black line), Tango (grey line) and Waltz (dotted black line). Since the raw scores obtained with the different softwares are not directly comparable, data were normalized to 100 as the highest score.

On-column refolding of denatured $Clb5^{\Delta156}$

Among different buffers used to solubilize the IBs, the best was a sodium phosphate/tris-based buffer at pH 12, with 8 M urea, although at high pH also several contaminant proteins were solubilized from IBs preparation [42]. Conventional methods for refolding insoluble recombinant proteins include slow dialysis or dilution into a large volume of refolding buffer or chromatographic refolding using packed columns [43]. Refolding using packed columns is attractive because it is easily automated using commercially available preparative chromatography systems. There are three basic approaches: (1) immobilization of the denatured protein onto a matrix and subsequent denaturant dilution to promote refolding; (2) denaturant dilution using size exclusion chromatography (SEC); and (3) immobilization of folding catalysts onto chromatographic supports so that the column behaves like a catalytic folding reactor [43]. Immobilization of denatured protein onto a matrix can be achieved through non-specific interactions or through specific affinity interactions. The aim is to isolate individual protein molecules spatially, thus inhibiting aggregation. Nickel-chelating chromatography has also been used for affinity immobilization and refolding [44-45].

We develop an on-column refolding procedure for $Clb5^{\Delta156}$applied to solubilized IBs loaded on Ni^{2+}/NTA resin. Urea concentration was suddenly

lowered by washing the resin with at least six column volumes of the sodium phosphate buffer containing different imidazole concentration. The protein was finally eluted with high-concentrated imidazole in sodium phosphate buffer. Fig 5 shows the purification results, indicating that the solubilized IBs bind to the resin in an efficient manner, indeed no bands corresponding to $Clb5^{\Delta 156}$ are detectable in the lanes containing the flow-through and washing step samples. A protein of the expected molecular weight for $Clb5^{\Delta 156}$ can be eluted from the resin in a native buffer, strongly suggesting that the protein is able to re-acquire the proper tertiary structure to remain soluble. This procedure allows the recovery of soluble $Clb5^{\Delta 156}$ with quite low yields (~ 10%) , and further improvements will be required to purify higher amount of protein to perform biochemical and biophysical analyses.

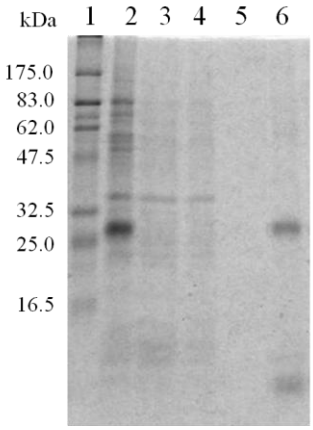


Figure 5. On-column refolding of $Clb5^{\Delta 156}$ analyzed by Coomassie-stained SDS-PAGE gels. Lane 1: molecular-weight markers; lane 2: solubilized IBs (sample loaded on IMAC) ; lane 3: not bound; lane 4: washing step 1; lane 5: washing step 2; lane 6: IMAC-eluted protein. For each lane, 150 μL of trichloroacetic acid (TCA)-precipitated sample were loaded. Initial volumes were as follows: solubilized IBs: 15 mL; not bound: 15 mL; washing step 1: 1 mL; washing step 2: 1 mL; IMAC-eluted protein: 500 μL.

Conclusions

The here reported data show the first results concerning the heterologous expression and the purification of Cdk1 and Clb5 from *S. cerevisiae*. Cdk1 is expressed at quite high levels, and is found both in the soluble and insoluble fractions of the intracellular proteins. The soluble fraction can be extracted by treating the cells with a gentle lysis buffer. CD analysis allowed us to

verify that the protein acquires a fold comparable to that reported for orthologous native kinase, hence it is likely that Cdk1 in this form can carry out its biological function and the ability to bind its physiological partners. On the contrary, full-length Clb5 undergoes to proteolytic degradation when expressed in *E. coli*. This behavior is likely due to the highly-disordered N-terminal region of the protein. The deletion of the first 156 amino acids allowed the protein to be efficiently overexpressed as IBs, due to the high aggregation propensity of the truncated protein $Clb5^{\Delta 156}$. A purification procedure was then developed to solubilize IBs and to refold $Clb5^{\Delta 156}$ by an on-column procedure.

These preliminary results can represent encouraging starting points for future improvements. The production of a functional Clb5-Cdk1 complex is essential for the complete description of the inhibition by Sic1. Indeed, several data have been presented on the kinase-inhibitor domain of Sic1, but they only concern the protein in the unbound state.

Therefore, the expression and purification of Clb5-Cdk1 complex is fundamental to elucidate more in detail the molecular mechanisms regulating the G1/S transitions in the *S. cerevisiae* cell cycle.

References

1. Mendenhall, M.D. and A.E. Hodge, *Regulation of Cdc28 cyclin-dependent protein kinase activity during the cell cycle of the yeast Saccharomyces cerevisiae.* Microbiol Mol Biol Rev, 1998. **62**(4): p. 1191-243.
2. Schwob, E., et al., *The B-type cyclin kinase inhibitor p40SIC1 controls the G1 to S transition in S. cerevisiae.* Cell, 1994. **79**(2): p. 233-44.
3. Nash, P., et al., *Multisite phosphorylation of a CDK inhibitor sets a threshold for the onset of DNA replication.* Nature, 2001. **414**(6863): p. 514-21.
4. Deshaies, R.J. and J.E. Ferrell, Jr., *Multisite phosphorylation and the countdown to S phase.* Cell, 2001. **107**(7): p. 819-22.
5. Verma, R., et al., *Phosphorylation of Sic1p by G1 Cdk required for its degradation and entry into S phase.* Science, 1997. **278**(5337): p. 455-60.
6. Brocca, S., et al., *Order propensity of an intrinsically disordered protein, the cyclin-dependent-kinase inhibitor Sic1.* Proteins, 2009. **76**(3): p. 731-46.
7. Hodge, A. and M. Mendenhall, *The cyclin-dependent kinase inhibitory domain of the yeast Sic1 protein is contained within the C-terminal 70 amino acids.* Mol Gen Genet, 1999. **262**(1): p. 55-64.
8. Brocca, S., et al., *Compaction properties of an intrinsically disordered protein: Sic1 and its kinase-inhibitor domain.* Biophys J, 2011. **100**(9): p. 2243-52.

9. Lambrughi, M., et al., *Intramolecular interactions stabilizing compact conformations of the intrinsically disordered kinase-inhibitor domain of Sic1: a molecular dynamics investigation.* Front Physiol, 2012. **3**: p. 435.
10. Mittag, T., et al., *Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor.* Proc Natl Acad Sci U S A, 2008. **105**(46): p. 17772-7.
11. Tang, X., et al., *Composite low affinity interactions dictate recognition of the cyclin-dependent kinase inhibitor Sic1 by the SCFCdc4 ubiquitin ligase.* Proc Natl Acad Sci U S A, 2012. **109**(9): p. 3287-92.
12. Mitrea, D.M., et al., *Disorder-function relationships for the cell cycle regulatory proteins p21 and p27.* Biol Chem, 2012. **393**(4): p. 259-74.
13. Peter, M. and I. Herskowitz, *Joining the complex: cyclin-dependent kinase inhibitory proteins and the cell cycle.* Cell, 1994. **79**(2): p. 181-4.
14. Barberis, M., et al., *The yeast cyclin-dependent kinase inhibitor Sic1 and mammalian p27Kip1 are functional homologues with a structurally conserved inhibitory domain.* Biochem J, 2005. **387**(Pt 3): p. 639-47.
15. Galea, C.A., et al., *Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits.* Biochemistry, 2008. **47**(29): p. 7598-609.
16. Lacy, E.R., et al., *p27 binds cyclin-CDK complexes through a sequential mechanism involving binding-induced protein folding.* Nat Struct Mol Biol, 2004. **11**(4): p. 358-64.
17. Sivakolundu, S.G., D. Bashford, and R.W. Kriwacki, *Disordered p27Kip1 exhibits intrinsic structure resembling the Cdk2/cyclin A-bound conformation.* J Mol Biol, 2005. **353**(5): p. 1118-28.
18. Russo, A.A., et al., *Crystal structure of the p27Kip1 cyclin-dependent-kinase inhibitor bound to the cyclin A-Cdk2 complex.* Nature, 1996. **382**(6589): p. 325-31.
19. Bowman, P., et al., *Thermodynamic characterization of interactions between p27(Kip1) and activated and non-activated Cdk2: intrinsically unstructured proteins as thermodynamic tethers.* Biochim Biophys Acta, 2006. **1764**(2): p. 182-9.
20. Alberghina, L., P. Coccetti, and I. Orlandi, *Systems biology of the cell cycle of Saccharomyces cerevisiae: From network mining to system-level properties.* Biotechnol Adv, 2009. **27**(6): p. 960-78.
21. Alfieri, R., et al., *Towards a systems biology approach to mammalian cell cycle: modeling the entrance into S phase of quiescent fibroblasts after serum stimulation.* BMC Bioinformatics, 2009. **10 Suppl 12**: p. S16.
22. Borriello, A., et al., *Targeting p27Kip1 protein: its relevance in the therapy of human cancer.* Expert Opin Ther Targets, 2011. **15**(6): p. 677-93.
23. Ryan, D.P. and J.M. Matthews, *Protein-protein interactions in human disease.* Curr Opin Struct Biol, 2005. **15**(4): p. 441-6.
24. Dobson, C.M., *Protein folding and misfolding.* Nature, 2003. **426**(6968): p. 884-90.
25. Chiti, F. and C.M. Dobson, *Protein misfolding, functional amyloid, and human disease.* Annu Rev Biochem, 2006. **75**: p. 333-66.
26. Obradovic, Z., et al., *Predicting intrinsic disorder from amino acid sequence.* Proteins, 2003. **53 Suppl 6**: p. 566-72.

27. Romero, P., et al., *Sequence complexity of disordered protein.* Proteins, 2001. **42**(1): p. 38-48.
28. Li, X., et al., *Predicting Protein Disorder for N-, C-, and Internal Regions.* Genome Inform Ser Workshop Genome Inform, 1999. **10**: p. 30-40.
29. Garner, E., et al., *Predicting Binding Regions within Disordered Proteins.* Genome Inform Ser Workshop Genome Inform, 1999. **10**: p. 41-50.
30. Sambrook, J., E.F. Fritsch, and T. Maniatis, *Molecular cloning. A laboratory manual.* Cold Spring Harbor, NY, 1989. **Cold Spring Harbor Laboratory Press.**
31. Matsumura, I. and L.A. Rowe, *Whole plasmid mutagenic PCR for directed protein evolution.* Biomol Eng, 2005. **22**(1-3): p. 73-9.
32. Hanahan, D., *Techniques for transformation of E. coli.* In D. M. Glover (Ed.), *DNA cloning: A practical approach.* (p. 109), 1985. **Mc Lean, VA: IRL press.**
33. *Promega Subcloning Notebook 44.* http://www.promega.com./guides/subcloning_guide/.
34. Studier, F.W., *Protein production by auto-induction in high density shaking cultures.* Protein Expr Purif, 2005. **41**(1): p. 207-34.
35. Laemmli, U.K., *Cleavage of structural proteins during the assembly of the head of bacteriophage T4.* Nature, 1970. **227**(5259): p. 680-5.
36. Sreerama, N. and R.W. Woody, *Estimation of protein secondary structure from circular dichroism spectra: comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set.* Anal Biochem, 2000. **287**(2): p. 252-60.
37. Rodrigues, A.J., et al., *Absence of ataxin-3 leads to cytoskeletal disorganization and increased cell death.* Biochim Biophys Acta, 2010. **1803**(10): p. 1154-63.
38. Child, E.S., et al., *A cancer-derived mutation in the PSTAIRE helix of cyclin-dependent kinase 2 alters the stability of cyclin binding.* Biochim Biophys Acta, 2010. **1803**(7): p. 858-64.
39. Xie, H., et al., *Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions.* J Proteome Res, 2007. **6**(5): p. 1882-98.
40. Brown, N.R., et al., *The crystal structure of cyclin A.* Structure, 1995. **3**(11): p. 1235-47.
41. Hartley, D.L. and J.F. Kane, *Properties of inclusion bodies from recombinant Escherichia coli.* Biochem Soc Trans, 1988. **16**(2): p. 101-2.
42. Singh, S.M. and A.K. Panda, *Solubilization and refolding of bacterial inclusion body proteins.* J Biosci Bioeng, 2005. **99**(4): p. 303-10.
43. Middelberg, A.P., *Preparative protein refolding.* Trends Biotechnol, 2002. **20**(10): p. 437-43.
44. Rogl, H., et al., *Refolding of Escherichia coli produced membrane protein inclusion bodies immobilised by nickel chelating chromatography.* FEBS Lett, 1998. **432**(1-2): p. 21-6.
45. Zahn, R., C. von Schroetter, and K. Wuthrich, *Human prion proteins expressed in Escherichia coli and purified by high-affinity column refolding.* FEBS Lett, 1997. **417**(3): p. 400-4.

3. REFERENCES

Abragam A (1994) Oxford University Press
Akke M (2002) Curr Opin Struct Biol *12*, 642-7.
Amadei A *et al.* (1993) Proteins *17*, 412-25.
Arrondo JL and Goni FM (1999) Prog Biophys Mol Biol *72*, 367-405.
Barberis M *et al.* (2005a) Biochem J *387*, 639-47.
Barberis M *et al.* (2005b) Biochem Biophys Res Commun *336*, 1040-8.
Barth A (2007) Biochim Biophys Acta *1767*, 1073-101.
Barth A and Zscherp C (2002) Q Rev Biophys *35*, 369-430.
Baumketner A *et al.* (2006) Protein Sci *15*, 420-8.
Benesch JL *et al.* (2007) Chem Rev *107*, 3544-67.
Bernstein SL *et al.* (2004) J Am Soc Mass Spectrom *15*, 1435-43.
Bienkiewicz EA *et al.* (2002) Biochemistry *41*, 752-9.
Borysik AJ *et al.* (2004) J Biol Chem *279*, 27069-77.
Brocca S *et al.* (2009) Proteins *76*, 731-46.
Brown NR *et al.* (1995) Structure *3*, 1235-47.
Brown NR *et al.* (1999) Nat Cell Biol *1*, 438-43.
Cai S and Singh BR (2004) Biochemistry *43*, 2541-9.
Careri M *et al.* (2004) Eur J Mass Spectrom (Chichester, Eng) *10*, 429-36.
Chatterjee A *et al.* (2005) J Chem Sci *117*, 3-21.
Chowdhury SK and Chait BT (1990) Biochem Biophys Res Commun *173*, 927-31.
Chowdhury SK *et al.* (1990) J Am Chem Soc *112*, 9012-9013.
Clegg GA and Dole M (1971) Biopolymers *10*, 821-6.
Cohen MJ and Karasek FW (1970) J Chromatogr Sci *8*, 330-337.
Corbett RJ and Roche RS (1984) Biochemistry *23*, 1888-94.
Cortese MS *et al.* (2005) J Proteome Res *4*, 1610-8.
Cox CJ *et al.* (2002) FEBS Lett *527*, 303-8.
Csizmok V *et al.* (2005) Biochemistry *44*, 3955-64.
Day PJ *et al.* (2009) Proc Natl Acad Sci U S A *106*, 4166-70.
De Bondt HL *et al.* (1993) Nature *363*, 595-602.
de Bruin RA *et al.* (2004) Cell *117*, 887-98.
Demchenko AP (2001) J Mol Recognit *14*, 42-61.
Dobo A and Kaltashov IA (2001) Anal Chem *73*, 4763-73.
Dosztanyi Z *et al.* (2009) Bioinformatics *25*, 2745-6.
Dosztanyi Z and Tompa P (2008) Methods Mol Biol *426*, 103-15.
Dunker AK *et al.* (2002) Biochemistry *41*, 6573-82.
Dunker AK *et al.* (2001) J Mol Graph Model *19*, 26-59.
Dunker AK *et al.* (2008a) BMC Genomics *9 Suppl 2*, S1.
Dunker AK *et al.* (2008b) Curr Opin Struct Biol *18*, 756-64.
Dunker AK and Uversky VN (2010) Curr Opin Pharmacol *10*, 782-8.
Dyson HJ and Wright PE (1998) Nat Struct Biol *5 Suppl*, 499-503.
Escote X *et al.* (2004) Nat Cell Biol *6*, 997-1002.
Evans T *et al.* (1983) Cell *33*, 389-96.
Fenn LS and McLean JA (2008) Anal Bioanal Chem *391*, 905-9.
Fligge TA *et al.* (1998) J Chromatogr B Biomed Sci Appl *706*, 91-100.
Fuxreiter M *et al.* (2004) J Mol Biol *338*, 1015-26.
Galea CA *et al.* (2008) Biochemistry *47*, 7598-609.
Gardner KH and Kay LE (1998) Annu Rev Biophys Biomol Struct *27*, 357-406.

Goormaghtigh E *et al.* (1999) Biochim Biophys Acta *1422*, 105-85.
Grandori R (2003) Curr Org Chem *7*, 1589-1603.
Grandori R *et al.* (2009) Biotechnol J *4*, 73-87.
Greenfield NJ (1999) Trends Analyt Chem *18*, 236-244.
Greenfield NJ (2004) Methods Mol Biol *261*, 55-78.
Griffin JH *et al.* (1972) J Biol Chem *247*, 6482-90.
Gunasekaran K *et al.* (2003) Trends Biochem Sci *28*, 81-5.
Harper JW and Elledge SJ (1998) Genes Dev *12*, 285-9.
Harper JW *et al.* (1995) Mol Biol Cell *6*, 387-400.
Heitz F *et al.* (1997) Biochemistry *36*, 4995-5003.
Henderson SC *et al.* (1999) Anal Chem *71*, 291-301.
Henzler-Wildman K and Kern D (2007) Nature *450*, 964-72.
Hess B (2000) Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics *62*, 8438-48.
Hodge A and Mendenhall M (1999) Mol Gen Genet *262*, 55-64.
Huang A and Stultz CM (2008) PLoS Comput Biol *4*, e1000155.
Hubbard SJ *et al.* (1998) Protein Eng *11*, 349-59.
Iakoucheva LM *et al.* (2002) J Mol Biol *323*, 573-84.
Iakoucheva LM *et al.* (2001) Protein Sci *10*, 1353-62.
Iakoucheva LM *et al.* (2004) Nucleic Acids Res *32*, 1037-49.
Invernizzi G *et al.* (2006) J Mass Spectrom *41*, 717-27.
Iribarne JV and Thomson BA (1976) J Chem Phys *64*, 2287-2294.
Jeffrey PD *et al.* (1995) Nature *376*, 313-20.
Kaddis CS and Loo JA (2007) Anal Chem *79*, 1778-84.
Kaltashov IA and Abzalimov RR (2008) J Am Soc Mass Spectrom *19*, 1239-46.
Kaltashov IA and Eyles SJ (2002) Mass Spectrom Rev *21*, 37-71.
Kaltashov IA and Mohimen A (2005) Anal Chem *77*, 5370-9.
Kamura T *et al.* (2003) Proc Natl Acad Sci U S A *100*, 10231-6.
Kanu AB *et al.* (2008) J Mass Spectrom *43*, 1-22.
Karplus M and McCammon JA (2002) Nat Struct Biol *9*, 646-52.
Kebarle P and Verkerk UH (2009) Mass Spectrom Rev *28*, 898-917.
Kelly SM *et al.* (2005) Biochim Biophys Acta *1751*, 119-39.
Kelly SM and Price NC (2000) Curr Protein Pept Sci *1*, 349-84.
Konermann L and Douglas DJ (1997) Biochemistry *36*, 12296-302.
Konermann L and Douglas DJ (1998) Rapid Commun Mass Spectrom *12*, 435-42.
Krell T *et al.* (1996) J Biol Chem *271*, 24492-7.
Krittanai C and Johnson WC (1997) Anal Biochem *253*, 57-64.
Kriwacki RW *et al.* (1996) Proc Natl Acad Sci U S A *93*, 11504-9.
Kuprowski MC and Konermann L (2007) Anal Chem *79*, 2499-506.
Kussie PH *et al.* (1996) Science *274*, 948-53.
Lacy ER *et al.* (2004) Nat Struct Mol Biol *11*, 358-64.
Lacy ER *et al.* (2005) J Mol Biol *349*, 764-73.
Lei H *et al.* (2007) Proc Natl Acad Sci U S A *104*, 4925-30.
Lemieux RU and Spohr U (1994) Adv Carbohydr Chem Biochem *50*, 1-20.
Lengronne A and Schwob E (2002) Mol Cell *9*, 1067-78.
Li X *et al.* (1999) Genome Inform Ser Workshop Genome Inform *10*, 30-40.
Liu J and Kipreos ET (2000) Mol Biol Evol *17*, 1061-74.
MacKerell AD *et al.* (1998) J Phys Chem *102*, 3586-3616.
Mann M and Wilm M (1995) Trends Biochem Sci *20*, 219-24.

Marsh JA *et al.* (2010) Structure *18*, 1094-103.
Martin SR and Schilstra MJ (2008) Methods Cell Biol *84*, 263-93.
Mendenhall MD and Hodge AE (1998) Microbiol Mol Biol Rev *62*, 1191-243.
Mirza UA and Chait BT (1994) Anal Chem *66*, 2898-904.
Mitrea DM *et al.* (2012) Biol Chem *393*, 259-74.
Mittag T *et al.* (2008) Proc Natl Acad Sci U S A *105*, 17772-7.
Mora JF *et al.* (2000) J Mass Spectrom *35*, 939-52.
Morgan DO (2007) Oxford University Press, 297 p.
Morgan DO (1997) Annu Rev Cell Dev Biol *13*, 261-91.
Mulder FAA *et al.* (2010) John Wiley & Sons, Inc.
Nash P *et al.* (2001) Nature *414*, 514-21.
Natalello A *et al.* (2005) Biochem J *385*, 511-7.
Natalello A *et al.* (2007) Biochemistry *46*, 543-53.
Nugroho TT and Mendenhall MD (1994) Mol Cell Biol *14*, 3320-8.
Nurse P (1990) Nature *344*, 503-8.
Olashaw N *et al.* (2004) Cell Cycle *3*, 263-4.
Oldfield CJ *et al.* (2005) Biochemistry *44*, 12454-70.
Ou L *et al.* (2011) J Biol Chem *286*, 30142-51.
Papaleo E *et al.* (2009) J Mol Graph Model *27*, 889-99.
Patriksson A *et al.* (2007) Biochemistry *46*, 933-45.
Pavletich NP (1999) J Mol Biol *287*, 821-8.
Pentony MM and Jones DT (2010) Proteins *78*, 212-21.
Peter M and Herskowitz I (1994) Cell *79*, 181-4.
Porath J (1968) Nature *218*, 834-8.
Prakash S *et al.* (2004) Nat Struct Mol Biol *11*, 830-7.
Price NC and Stevens E (1983) Biochem J *213*, 595-602.
Ptitsyn OB (1995) Adv Protein Chem *47*, 83-229.
Puntervoll P *et al.* (2003) Nucleic Acids Res *31*, 3625-30.
Purves RW and Guevremont R (1999) Anal Chem *71*, 2346-57.
Radivojac P *et al.* (2007) Biophys J *92*, 1439-56.
Reeves R and Nissen MS (1999) Methods Enzymol *304*, 155-88.
Romero PR *et al.* (2006) Proc Natl Acad Sci U S A *103*, 8390-5.
Russell P *et al.* (1989) Cell *57*, 295-303.
Russo AA *et al.* (1996a) Nature *382*, 325-31.
Russo AA *et al.* (1996b) Nat Struct Biol *3*, 696-700.
Russo AA *et al.* (1998) Nature *395*, 237-43.
Sanglier S *et al.* (2008) Methods Mol Biol *484*, 217-43.
Schneider BB *et al.* (2001) Rapid Commun Mass Spectrom *15*, 249-57.
Schwob E *et al.* (1994) Cell *79*, 233-44.
Schwob E and Nasmyth K (1993) Genes Dev *7*, 1160-75.
Sherr CJ and Roberts JM (1999) Genes Dev *13*, 1501-12.
Shoemaker BA *et al.* (2000) Proc Natl Acad Sci U S A *97*, 8868-73.
Sickmeier M *et al.* (2007) Nucleic Acids Res *35*, D786-93.
Sigalov AB *et al.* (2007) Biochimie *89*, 419-21.
Sivakolundu SG *et al.* (2005) J Mol Biol *353*, 1118-28.
Slep KC *et al.* (2001) Nature *409*, 1071-7.
Sreerama N and Woody RW (2000) Anal Biochem *287*, 252-60.
Strickland EH (1974) CRC Crit Rev Biochem *2*, 113-75.
Sugase K *et al.* (2007) Nature *447*, 1021-5.

Surewicz WK *et al.* (1993) Biochemistry *32*, 389-94.
Takaki T *et al.* (2009) Proc Natl Acad Sci U S A *106*, 4171-6.
Tompa P (2005) FEBS Lett *579*, 3346-54.
Tompa P (2002) Trends Biochem Sci *27*, 527-33.
Tompa P and Csermely P (2004) FASEB J *18*, 1169-75.
Tompa P and Fuxreiter M (2008) Trends Biochem Sci *33*, 2-8.
Tompa P *et al.* (2009) Bioessays *31*, 328-35.
Tompa P *et al.* (2005) Trends Biochem Sci *30*, 484-9.
Toyn JH *et al.* (1997) Genetics *145*, 85-96.
Tyers M *et al.* (1993) EMBO J *12*, 1955-68.
Ubersax JA *et al.* (2003) Nature *425*, 859-64.
Uhlmann F (2003) Biochem Soc Symp, 243-51.
Uversky VN (1994) Int J Bio-Chromatogr *1*, 103-14.
Uversky VN (2011) Biochim Biophys Acta *1814*, 693-712.
Uversky VN (2010) J Biomed Biotechnol *2010*, 568068.
Uversky VN (2002a) Protein Sci *11*, 739-56.
Uversky VN (2003) Cell Mol Life Sci *60*, 1852-71.
Uversky VN (2002b) Eur J Biochem *269*, 2-12.
Uversky VN and Dunker AK (2010) Biochim Biophys Acta *1804*, 1231-64.
Uversky VN *et al.* (2008) Annu Rev Biophys *37*, 215-46.
Uversky VN *et al.* (2009) BMC Genomics *10 Suppl 1*, S7.
Vacic V *et al.* (2007) BMC Bioinformatics *8*, 211.
Venyaminov S and Prendergast FG (1997) Anal Biochem *248*, 234-45.
Verkerk UH and Kebarle P (2005) J Am Soc Mass Spectrom *16*, 1325-41.
Vorobjev YN *et al.* (1998) Proteins *32*, 399-413.
Wallace BA (2000) J Synchrotron Radiat *7*, 289-95.
Wang Y *et al.* (2005) Chembiochem *6*, 2242-6.
Wang Y *et al.* (2011) Nat Chem Biol *7*, 214-21.
Ward JJ *et al.* (2004) J Mol Biol *337*, 635-45.
Willems AR *et al.* (2004) Biochim Biophys Acta *1695*, 133-70.
Wuthrich K (1986) Wiley.
Wuthrich K (1990) J Biol Chem *265*, 22059-62.
Xie H *et al.* (2007) J Proteome Res *6*, 1882-98.

ACKNOWLEDGEMENTS

I would like to thank…vabbè direi che posso chiuderla qui con l'inglese.

Grazie alla Prof. ssa Grandori, per avermi dato la possibilità di lavorare con lei, nonostante tutto quello che le ho combinato, e alla Prof. ssa Lotti, per avermi concesso di stare nel suo laboratorio, da infiltrato, per tutto questo tempo.
Un più-che-doveroso ringraziamento va alla Stefy, per tutto quello che ha fatto, ma soprattutto per tutto quello che è stata, per me…

Grazissime a tutti i ragazzi del lab, anzi dei labs, presenti (Giusy, Fede, Fra, Carlo, Antonino, Mattia, Susy, Simo1, Simo2, Manuela), passati (Marco, che rimarrà sempre nel mio cuuore, Chiara, Stefano, Ila, Pietro, Maia, e credo più o meno altri cento che la mia memoria da elefante mi hanno fatto dimenticare), futuri. Serberò un fantastico ricordo di tutti quanti, soprattutto di quelli futuri.
Grazie a Nadia che mi ha portato a stampare la tesi (anche se, nel momento in cui sto scrivendo i ringraziamenti, questo non è ancora successo).

Grazie alla mia famiglia, e a Chiara - che ormai è come dire la stessa cosa praticamente-.