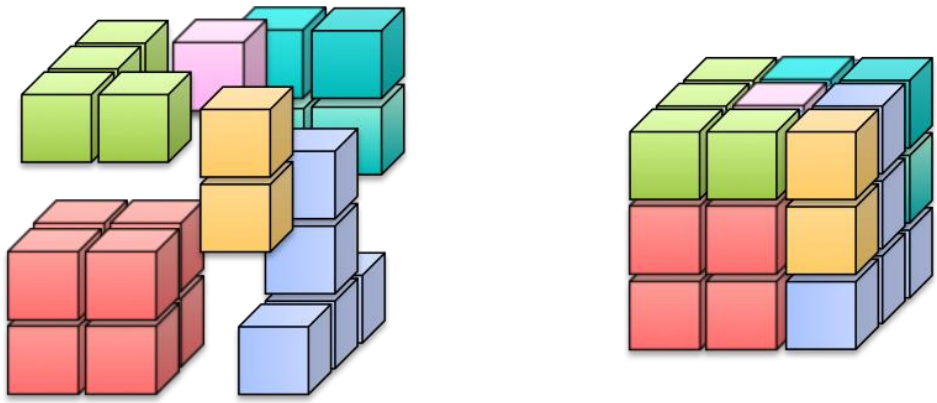
SAPIENZA
UNIVERSITÀ DI ROMA

PhD Course in Biochemistry

XXXV Cycle (Academic Years 2019-2022)

Comparison of the folding and the binding features of domains in isolation and in a multidomain construct



PhD student
Livia Pagano

Tutor
Prof. Stefano Gianni

PhD coordinator
Prof. Stefano Gianni

December 2022

1. Introduction

1.1 The protein folding problem

Using a literary metaphor, we might consider the twenty natural amino acids characterizing the biological active proteome, as the letters that compose a whole speaking language. The discrepancy between the limited numbers of units needed to build active proteins and the variety of all the vital activities generated by these structures, is the baseline of the protein folding problem: how the particular polypeptide chain holds the information of its own active tridimensional structure? [1,2]

One of the milestones of the protein folding field is the work of Christian Anfinsen on the (un)folding of Ribonuclease A, for which he has been awarded with the Nobel Prize in the 1972[3]. Now, 50 years later, DeepMind company has released an artificial intelligence system that can predict protein 3D structures with accuracy comparable to the experiments[4]. Big steps have been made to solve the protein folding problem, but still the relationship between the amino acids sequence of a protein and its biologically active native state needs to be understood.

Protein folding is cooperative

Just a few years after Anfinsen's work on the reversibility of the chemical denaturation of Ribonuclease A, studies have been carried out on the mechanism of folding of proteins in solution. In particular, the biochemist Charles Tanford described the transition between the native and denaturate state of a protein, as a cooperative phenomenon[5]. As shown in Figure 1. small amounts of denaturant don't affect the whole structure of the protein, which can resist the destabilizing environment. At higher denaturant concentration an all-or-none transition (represented mathematically by a

sigmoidal function) yields the protein to an unfolded state, in which the weak but numerous forces that contribute to the global stability have been broken cooperatively.

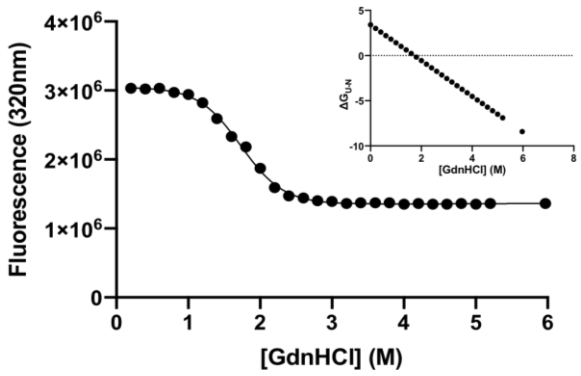


Figure 1. *Unfolding profile of the Grb2 SH2 domain. The transition is followed by measuring the change in fluorescence upon denaturant concentration. Inset panel shows the linear free energy extrapolation. As discussed in 1.2.1 paragraph, this analysis allows the estimation of the protein stability in absence of denaturant.*

Protein folding is not a random search

Evolution solved the folding problem, choosing specific pathways instead of random exploration, but describing how these pathways are been chosen is the major challenge in this field.

In the 1968 Cyrus Levinthal published "Are there pathways for protein folding?"[6], a question resulting from the mathematical paradox of the estimation of the folding rate of a protein. Even for a short peptide (e.g. 100 amino acids) the random inspection of all possible configurations would take more than the age of the universe, contrasting with the natural folding time scale of seconds or less. Therefore, there must be some constraints that lead

to the native state. A well-established answer to Levinthal's paradox is that the main interactions characterizing the native state, shape the energy landscape of a foldable protein[7–9]; in fact among all the possible conformations, proteins are minimally frustrated, therefore, there is a bias towards the native state that can be considered the bottom of a rugged funnel[1,10,11] (see Fig. 2).

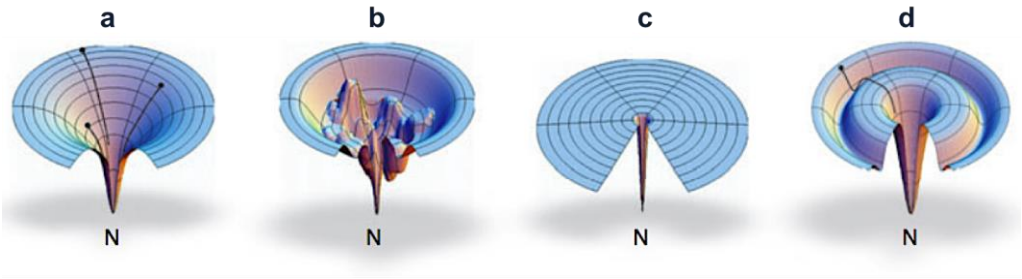


Figure 2. *Schematic tridimensional representation of the different shapes of funneled energy landscapes. (a) A smooth energy landscape for a fast folder, (b) a rugged energy landscape with kinetic traps, (c) a golf course energy landscape in which folding is dominated by diffusional conformational search, and (d) a moat landscape, where folding must pass through an obligatory intermediate. (Adapted from Dill et al 2008)*

1.1.1 Folding studies on multidomain constructs

Since Anfinsen's experiment on the single-domain protein RNase A, most of the folding studies have traditionally focused on domains in isolation. The reason of such scaling down is not just that larger proteins are tough to study, but because domains are always been considered structural and functional units that can fold independently[12] and, therefore, summed up to describe a bigger protein. Despite more than 70% of eukaryotic proteins are composed

by more than one domain[13–17], relatively few studies have been carried out on the folding multidomain proteins[18–26].

Over the last decades the folding community started to question whether the information gathered on the folding of a domain in isolation matches closely what should be observed in the context of more complex structural architectures; one of the first detailed review on the folding of domains in a larger-sized system was written by Jaenicke and coworkers (2000)[27], and a more comprehensive study was published by Batey et al[28].

Taking into account a two-domain construct, it is possible to summarize three main conditions:

1. The two units are independent of one another, resulting in a perfect sum of each kinetic and thermodynamic properties[29]
2. One domain facilitates the folding of the neighboring domain[30]
3. The presence of the second domain slows down the folding of the first domain, causing the accumulation of transient kinetic traps[20,21,24,31–36].

Characterizing the inter-domain interactions responsible for these effects is not trivial but there are experimental strategies that have been developed for such goal that will be discussed further on in this introduction.

1.2 Experimental strategies to study protein folding

How fast proteins fold? How much stable are the ground states? Studying *in vitro* the kinetics and the equilibrium of proteins is based on inducing progressive denaturation while monitoring an optical feature of the construct: fluorescence, absorbance, circular dichroism or nuclear magnetic resonance (NMR). The denaturation can be caused by changing temperature, pH, chemical environment[37] and, in the recent years, even applying a mechanical force as a mechanism to unfold the protein (AFM)[38]. Hence, measuring the

differences in the signals of those optical probes between the states that populates during the transition and the ground states enable to collect thermodynamic and kinetic information on the folding pathway of a protein.

1.2.1 Equilibrium studies

The experimental approach in order to assess the thermodynamic parameters of a protein at physiological condition is to insert a perturbation, e.g. adding chaotropic agents such as urea and guanidine, which can progressively affect the stability of the system. As discussed in the 1.1 paragraph, for a small single-domain protein, the transition from native to unfolded state is represented by a sigmoidal behavior, indicating a cooperative (un)folding process[37]. Plotting the calculated stability at different denaturant concentration results in a linear function, in which the intercept of the Y-axis referred to the protein stability in absence of denaturants (inset panel of Figure 1.). The slope of the line is an intrinsic constant of the protein, known as $m_{U\text{-}N}$ value that gives a degree of how the protein resists denaturation[39]. In addition, it has been shown empirically that the $m_{U\text{-}N}$ value correlates very strongly with the amount of protein surface exposed to solvent upon unfolding.

Thus, a system that displays a two-state transition at the equilibrium populates only the native and the unfolded states, as described in Scheme 1:

$$U \leftrightarrow D \qquad \text{(Scheme 1.)}$$

Taking into account the linear free energy correlation is possible to extrapolate the difference in free energy between the unfolded and the folded state:

$$\Delta G_{U-N} = \Delta G^{0}_{U-N} - m_{U-N}[\text{den}] \qquad \text{(Eq.1)}$$

The $\Delta G_{U\text{-}N}$, which is the stability of the protein at different denaturant concentrations, can also be derived using the mass action law as:

$$\Delta G_{U-N} = -\text{RT}\ln(K_{eq}) \quad \text{(Eq. 2)} \qquad \text{and} \qquad K_{eq} = [\text{U}]/[\text{N}] \qquad \text{(Eq. 3)}$$

Where K_{eq} is the equilibrium constant of the (un)folding reaction, R is the gas constant, T is the temperature, the [U] and [N] the concentration of the unfolded and native states respectively.

To test the robustness of the calculated thermodynamic parameters is recommended both to perform the equilibrium experiments monitoring different probes and to compare them with the ones derived from kinetics.

Unexpected m-value in multidomain systems

Although for a single two-state system the parameters resulting from this kind of analysis can be determined within a certain error, for a more complex system equilibrium denaturation can lead to different unfolding profiles. Considering Myers and coworkers's work[39], the m-value, thus, solvent accessible surface area, is related with the size of the protein; therefore for a larger size protein the m-value is expected to be higher than its own isolated domain.

Figure 3 describes two possible denaturation profiles of tandem repeats compared to the same domains expressed in isolation. In the case (a) the PDZ1-PDZ2 tandem of Whirlin protein has two midpoints significantly separated, therefore two transitions are observed[31]. Moreover, the double

sigmoid is perfectly consistent with the sum of the individual curves for PDZ1 and PDZ2 in isolation.

On the contrary, the unfolding denaturation shown in panel (b) the full-length protein displays a broader transition respect to the single domains. An apparent two-state transition hides the individual unfolding processes, leading to a lower m-value, consequently a lower total free energy of unfolding.

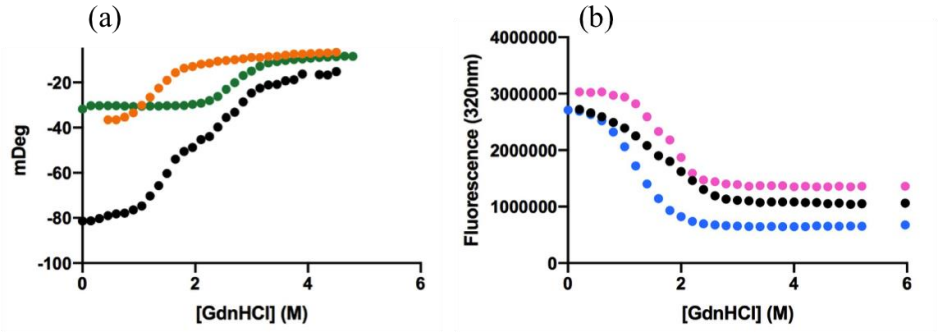


Figure 3. (a) *GdnHCl-induced equilibrium denaturation of PDZ1-PDZ2 (black circles), PDZ1 (orange circles), and PDZ2 (green circles) monitored by CD. The difference in the midpoints of the domains in the tandem allows the detection of two transitions. (b) GdnHCl-induced equilibrium denaturation of Grb2 protein (black circles), SH2 domain (pink circles) and the C-SH3 domain (blue circles). Only one transition can be observed in the multidomain protein, due to the closeness of the individual midpoints.*

1.2.2 Kinetic studies

While the thermodynamic studies provide information about the stability and the co-operativity of a protein, only the kinetics can describe the protein dynamics and the mechanical details of the folding pathway. Furthermore,

kinetics is essential to address the structural properties of folding transition states as well as in defining which amino acids are important for the folding of a given protein. A system described by a two-state transition in an equilibrium denaturation may be kinetically more complex, with a wrinkled folding pathway characterized by transiently populated intermediate(s).

A general kinetic experiment implies the rapid perturbation of the equilibrium by changing the chemical (mixing techniques) or physical (relaxation techniques) properties of the solution containing the protein of interest. Whenever the system is perturbed, the variation in an optically active property of the system is measured along time scale of the reaction.

This kind of analysis can be conducted using a stopped-flow device with pressure-driven syringes pushing the two solutions to a mixing chamber. The (un)folding reaction can be followed by monitoring the decrease (or increase) in the tryptophane's fluorescence intensity.

Two-state model

In a two-state reaction only the native and unfolded state accumulates, therefore it can be described with Scheme 2:

$$U \underset{k_u}{\overset{k_f}{\rightleftarrows}} N \qquad \text{(Scheme 2.)}$$

In these conditions, the stopped-flow traces are the best fit to a single exponential decay and the observed rate constant k_{obs} is:

$$k_{obs} = k_f^0 \exp\big(m_f[den]\big) + k_u^0 \exp\,(m_u[den]) \qquad \text{(Eq. 4)}$$

$$k_{obs} = k_f + k_u \qquad \text{(Eq. 5)}$$

where k_f^0 and k_u^0 are the folding and unfolding rate constants in the absence of denaturant, k_f and k_u the folding and unfolding rate constants, and m_f and m_u are the m-values of folding and unfolding[37].

A semilogarithmic plot of the observed rate constants against different denaturation concentrations, called a Chevron plot, is used to describe the kinetics of protein folding (see Figure 4). Kinetic analysis of a Chevron plot allows the determination of thermodynamic parameters of unfolding. In particular, for a two-state model:

$$\Delta G^0_{\text{U-N}} = -\text{RT}\ln(k_f/k_u) \qquad \text{(Eq.6)}$$

where $\Delta G^0_{\text{U-N}}$ is the stability of the protein in the absence of denaturant, k_f and k_u are the refolding and unfolding rate, R is the gas constant and T is the temperature. As discussed in the 1.2.1 paragraph, the m-value represents the difference in accessible surface area between the two states[39]. The total m_{U-N} value can be calculated as:

$$m_{U-N} = m_f + m_u \qquad \text{(Eq.7)}$$

where m_f and m_u are the slopes of the folding and unfolding (respectively) branches of the Chevron plots and can be used to estimate the position of the transition state along the folding reaction coordinate.

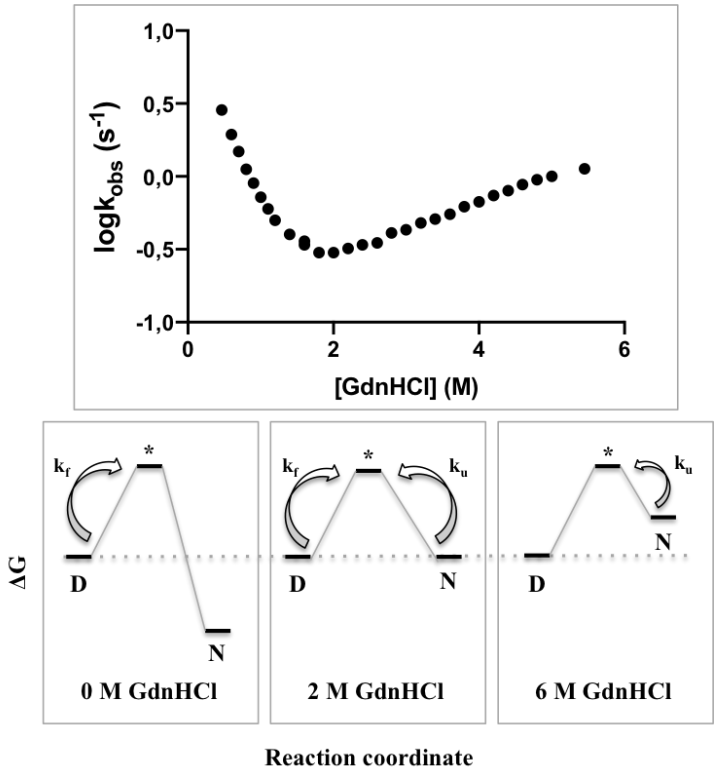


Figure 4. *Chevron plot of the C-terminal SH3 of Grb2 protein showing the dependence of the refolding and unfolding rate constants (k_f and k_u) upon GdnHCl. Below schematic representations of the folding free-energy profile at 0, 2 and 6 M GdnHCl. Empirically, the free-energy differences between D, *, and N changes linearly with [GdnHCl] resulting in a V-shaped chevron plot.*

The equilibrium and kinetic $\Delta G^0_{\text{U-N}}$ and m_{U-N} values should be robust and conserved at all the conditions explored; together with the linear dependence upon [GdnHCl] are crucial factors to verify the validity of a two-state model. Any deviation from these observations is a sign of a more complex folding model.

Three-state model

As discussed in 1.2.1 paragraph, many proteins that exhibit a two-state equilibrium transition, may have kinetics that is not two-state, involving the presence of transiently populated intermediates along the folding pathway[40]. When a folding intermediate is present, the observed rate constant becomes dependent on more than one energy barrier, and refolding is generally described by the sum of two or more exponential processes[41,42]. Usually in that case the resulted plot deviates from the classical V-shaped Chevron and a roll-over effect is observed (see Figure 5). Also in this case, the traces are the best fit to a single exponential decay but the observed rate constant k_{obs} is:

$$k_{obs} = \frac{k_f^0 \exp\left(-m_f[den]\right)}{(1+K_{eq}exp(m_{eq}[den])} + k_u^0 \exp\left(m_u[den]\right) \qquad \text{(Eq.8)}$$

where k_f^0 and k_u^0 are the folding and unfolding rate constants at zero molar denaturant concentration; $m_{f,}$ and m_u indicate their dependence from denaturant concentration, K_{eq} is the equilibrium constant between the denaturate and the intermediate and m_{eq} the associated m-value.

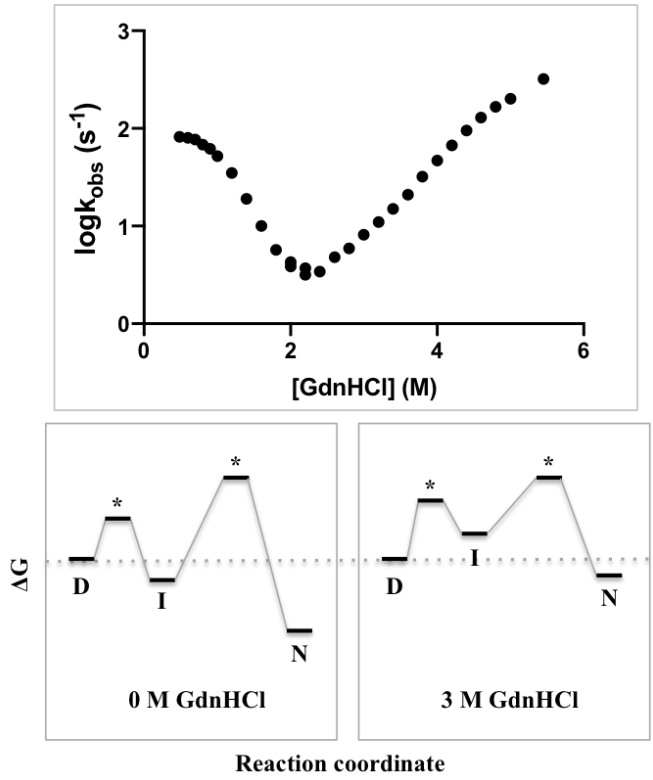


Figure 5. *The three-state chevron plot of the PDZ1 of Whirlin protein. Bottom panel the free-energy diagram at two representative guanidine concentrations, 0 M and 3 M. In the upper panel, at low denaturant concentration the refolding branch deviates from linearity, displaying slower folding rate constant compared to those expected from a two-state model. Increasing the denaturant concentration, the folding intermediate is less stable than the unfolded state, therefore the V-shape trend is conserved.*

1.2.3 Φ-value analysis: mapping transient states

The importance of the kinetic experiments is to provide detailed information about the actual folding pathway of a protein and, in addition, to structurally describe the transition(s) state(s) of the reaction. Generally, the folding reactions are highly cooperative, thus sometimes intermediates are not populated at the equilibrium. Moreover, by definition, transition states never accumulate (Eyring's Transition State Theory)[43] therefore studying the folding of these metastable states may be experimentally problematic. However, they can be observed indirectly combining protein engineering and folding rate constants analysis. Alan Fersht and co-workers developed a powerful technique to obtain information about transition states and intermediates of reactions called the Φ-value analysis[44]. This method consists in the introduction of a perturbation in the system by systematically mutating side chains and assessing their effect on the (un)folding rate constants. In fact, mutations that destabilize the intermediate target the contacts formed in the structure. In this way it is possible to map out interaction patterns of the intermediates and transition states. Schematic representation of the differences in free energy caused by mutations is reported in Figure 6.

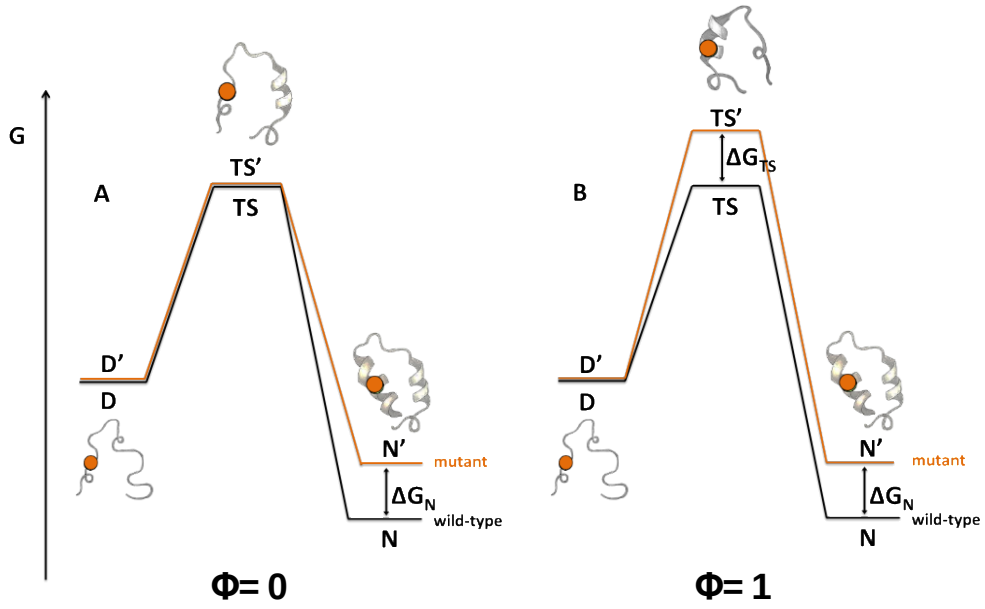


Figure 6. *Energy diagram of Φ-value analysis for a folding reaction. D, TS and N are the denaturated state, transition state and native state of the WT; D', TS' and N' are referred to the variant. The point mutation is represented as orange circle in all the structures.*

The Φ-value is calculated as the ratio between the change in the free energy variation between the denatured state (U) and the transition state (TS) and that between the unfolded state and the native state (N) upon mutation:

$$\Phi = \frac{\Delta\Delta G_{U-TS}}{\Delta\Delta G_{U-N}} \qquad \text{(Eq.9)}$$

Where the change in free energy variation between U and TS ($\Delta\Delta G_{U-TS}$) is defined as:

$$\Delta\Delta G_{U-TS} = \Delta G_{U-TS}^{WT} - \Delta G_{U-TS}^{MUT} = -RTln\frac{k_f^{WT}}{k_f^{MUT}} \qquad \text{(Eq.10)}$$

And the change in the total free energy variation ($\Delta\Delta G_{U-N}$) is defined as follow:

$$\Delta\Delta G_{U-N} = \Delta G_{U-N}^{WT} - \Delta G_{U-N}^{MUT} = RTlnK_{eq}^{MUT} - RTlnK_{eq}^{WT} \qquad \text{(Eq.11)}$$

$$\Delta\Delta G_{U-N} = RTln\frac{k_f^{MUT}k_u^{WT}}{k_u^{MUT}k_f^{WT}} \qquad \text{(Eq.12)}$$

K_{eq} is the equilibrium constant defined as the ratio between k_f and k_u. k_f and k_u are respectively the rate constant of folding and unfolding, experimentally obtained for the wt and the mutants through kinetic experiments. A Φ-value of 0 means that the residue is not perturb by the mutation, hence it is probably unfolded as the denaturated state and folds downhill the main barrier of folding; whereas a Φ-value of 1 indicates that the mutation has perturbed the transition state as the native state, meaning that the residue is probably structured in the transition state. Calculating the Φ-values for each variant allows mapping the structure of the metastable states, therefore to assess the degree of similarity with the native state and finally to hypothesize the mechanism through which folding occurs. It is the only experimental technique available for fine structural analysis of transition states of folding.

1.3 Studying protein binding *in vitro*

Most of the physiological cellular pathways, such as gene expression, cell growth, metabolism, proliferation and apoptosis, require protein-protein interactions (PPIs). For this reason, it is a major biological and biophysical

interest to understand how proteins associate and dissociate. Characterizing protein–protein interactions through methods such as NMR, X-Ray crystallography, fluorescence spectroscopy, pull-down assay, Co-immunoprecipitation, western blot, etc. is critical to understand protein function and the biology of the cell[45].

1.3.1 Equilibrium experiments

At a simplistic level, the non-covalent interaction between a protein (P) and a ligand (L) is often represented as follows:

$$P + L \underset{k_{off}}{\overset{k_{on}}{\rightleftarrows}} PL \qquad \text{(Scheme 3.)}$$

The term "ligand" in biological systems can represent different species. Usually, it is used to mean any molecule, which interacts with a given molecule (in this case a protein-protein complexes). At the equilibrium, the binding between P and L follow the law of mass action, and the dissociation constant is defined as:

$$K_d = \frac{[P][L]}{[PL]} = \frac{k_{off}}{k_{on}} \qquad \text{(Eq.13)}$$

K_d is a measure of the tendency of the PL complex to dissociate and k_{on} and k_{off} are the association and dissociation rate constants (dimensions concentration^{-1} time^{-1} and time^{-1} respectively).

In addition, the total concentration of protein P_{TOT} is conserved:

$$[P] + [PL] = [P]_{TOT} \qquad \text{(Eq.14)}$$

Combining equation 13 and equation 14 and solving the resulting equation for [PL] leads to the Langmuir isotherm (or hyperbola), which describes the binding between two single species:

$$[\mathrm{PL}] = \frac{[L][P]_{TOT}}{[L]+Kd} \qquad \text{(Eq.15)}$$

A simple equilibrium binding experiment consists in measuring the change in fluorescence of the tryptophan residues of a protein at increasing ligand concentration. As shown in Figure 7 a decrease in fluorescence is monitored upon increasing ligand concentration (the opposite trend can be observed) and the fit follows the hyperbolic function where the K_d is the ligand concentration, which half of the binding sites are occupied at equilibrium.

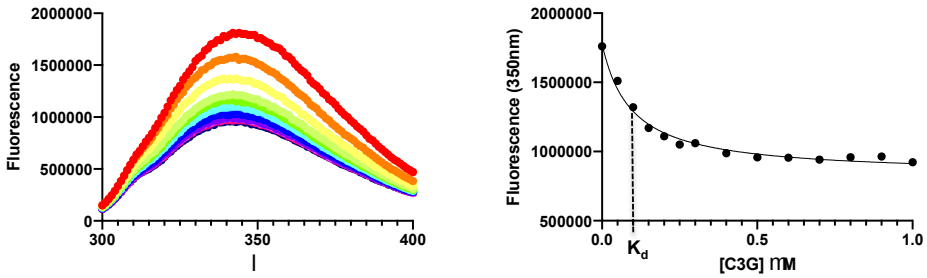


Figure 7. *Equilibrium binding experiment of the N-SH3 domain of CrkL with C3G ligand. Triptophane residue is excited at 280 nm and the emission is monitored between 300 and 400 nm. On the right panel the fluorescence measured at 350 nm is plotted versus ligand concentration and the data are fitted with equation 15.*

Knowing the K_d of a bimolecular complex can give information about the affinity between two reactants, however at cellular level, proteins often

engage in transient interactions i.e. the dissociation rate constant is in micromolar range, making them hard to be detected[46]. For this reason, different methods have been developed to identify low affinity, transient interactions, such as the phage display technique.

1.3.2 Phage display technique

A relatively recent tool to map and analyse protein-protein interactions is the proteomic peptide-phage display (Prop-PD), which is a variant of classical phage display developed by George P. Smith and Valery A. Petrenko in 1997[47,48] .

This high throughput technique takes advantage from the use of phages, viruses that infects bacterial cells. In fact, it is possible to genetically fuse all or part of a foreign protein to the exposed parts of the capsid without greatly impairing the phage infection cycle. In this way, the foreign amino acids are displayed at the tip of the virion, where they may be accessible to macromolecules like proteins. If a large collection of diverse DNAs – for example, DNAs representing all or most of the protein-encoding genes in an organism – are inserted into a phage DNA, the result is a highly diverse collection of phage chromosomes called an expression library. When the DNA library is transfected into bacterial cells, a peptide library is then expressed and displayed by the bacterial machinery. A very large number of variants (up to 10^{10}) can be selected in a single phage library, scaling up remarkably the search for novel protein binders.

Schematic representation of the phage display workflow is reported in Figure 8: after the phage library construction, the phages that bind the bait proteins are eluted and amplified several times. Peptide coding sequences of binding

enriched phages are then amplified using PCR before being sequenced through next-generation sequencing (NGS). NGS is relatively new technology for DNA and RNA analysis, which offers a simultaneous sequencing of thousands to millions of short nucleic acid sequences in a massive fashion[49]. Then the unique oligonucleotides are converted back to peptide sequences and matched back to the proteome library. A consensus binding motif is then generated using the enriched peptides sequences.

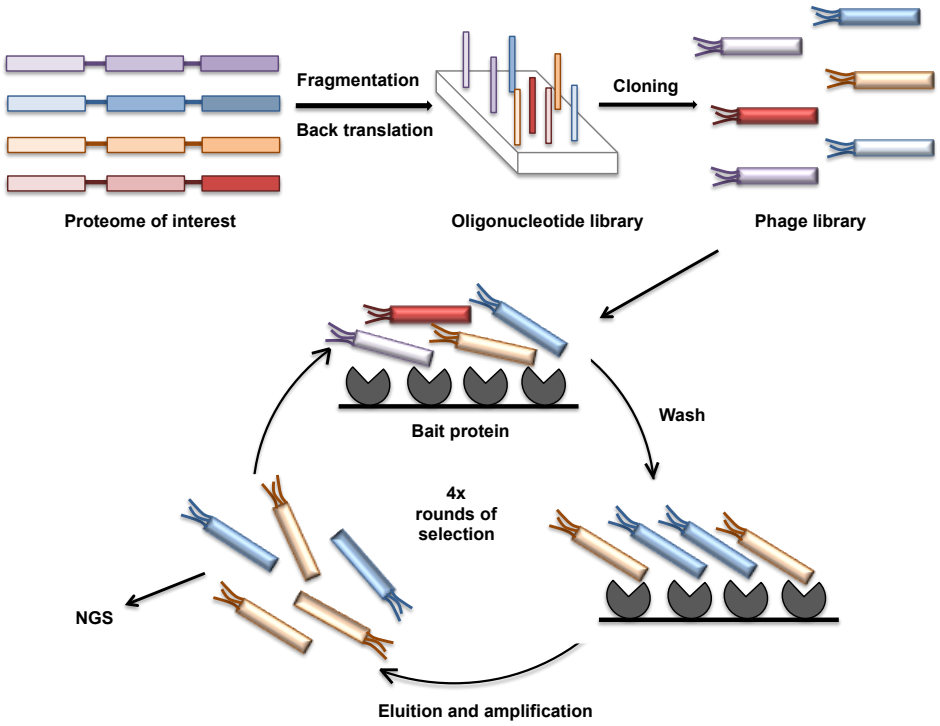


Figure 8. *Prop-PD workflow. Based on this method, the proteome of interest is segmented in short peptides. The DNA sequences of these peptides are cloned into a phagemid vector. The phage then infects the E. coli that produces the phage library. The library is then used to enrich phages that bind the target proteins. The selected phages are subjected to PCR and sequenced by NGS.*

1.4 Learning from two case studies: Whirlin and CrkL proteins

Recent studies have reported the importance of the inter-domain communication in the functionality of a protein. In fact, in both protein folding and binding, studying the effect of tethered domains within a more complex architecture is crucial for a better understanding of the physiological role of full-length proteins[64,65]. The effect of inter-domain interactions on the folding and the binding mechanisms is illustrated using two specific examples: Whirlin and CrkL proteins.

1.4.1 Whirlin: Does PDZ1 fold by the same pathway in isolation and in the PDZ1-PDZ2 tandem?

Whirlin is a large multi-domain scaffolding protein presents in both hair cells and retinal photoreceptor cells. In the inner ear it mechanically transmits, in complex with other proteins, the deflection of the stereocilia induced by sound waves. The full-length protein contains the N-terminal HHD1 (Harmonin Homology Domain), a tandem of PDZ domains (PDZ1-PDZ2), a HHD2, a large proline-rich region, and PDZ3 directly followed by a C-terminal PBM. One particular mutation responsible for Usher syndrome comes from the truncation of PDZ1-PDZ2 tandem (A207-K279 deletion), underlying the importance of this region in the function of the protein[50,51]. Delhommel and coworkers largely studied the structure of the tandem; their research reveals an atypical arrangement in which the PDZ domains are facing each other with a symmetric interaction of their binding site, leading to the burial of PDZ2 binding site, being hardly accessible for a ligand. The PDZ2 of whirlin is suggested to have evolved to favor PDZ1 stabilization, improving PDZ1 affinity by 3- to 4-fold[52,53].

The characterization of the folding kinetics of PDZ1-PDZ2 demonstrated the presence of a misfolded intermediate that competes with productive folding[31].

In particular, the fortuitous differences in stability between PDZ1 and PDZ2 allowed to show that only the concurrent denaturation of both domains leads to the accumulation of the kinetic trap that slows the productive pathway. Whereas when PDZ2 is held in its native conformation the refolding arm displays a detectable but less pronounced roll-over. Whilst it is clear that the presence of denatured PDZ2 has a pronounced effect on the folding of PDZ1, no studies have been carried out on the single domain when the former is held its native state.

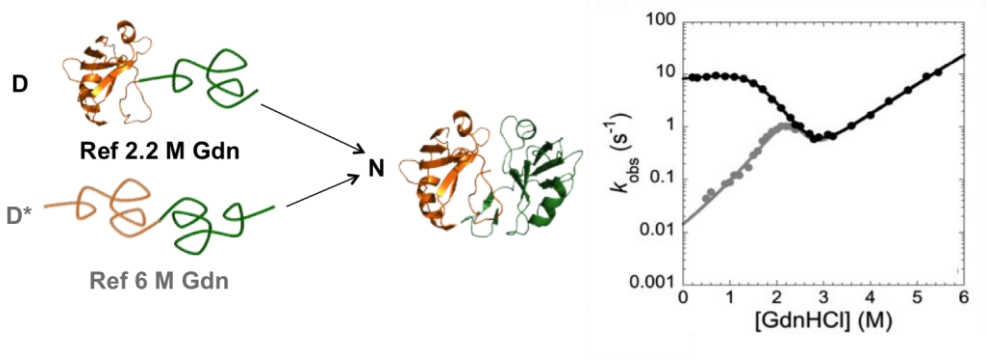


Figure 9. *On the left side of the panel is reported a schematic representation of the species involved in the refolding process. The PDZ2 is coloured in orange and the PDZ1 in green. As shown in the chevron plot, there are two populations that can be detected starting from mild or high denaturant concentration (i.e. 2.2 M or 6 M GdnHCl).*

1.4.2 CrkL: Are there any differences in the binding affinities between the N-terminal domain and the full-length protein?

CrkL is a ubiquitously expressed adapter protein, member of the proto-oncogene CRK family. It mediates and regulates several physiological pathways and it has been observed its prominent role in the onset of different kind of human cancers[54,55]. Intriguingly, CrkL possesses no catalytic or transcriptional activity and exerts its functions through its protein–protein

interaction modules that compose the entire protein, i.e., one N-terminal SH2 domain followed by two SH3 domains (namely N-SH3 and C-SH3). The SH2 domain binds short tyrosine phosphorylated proteins (pTyr-Asp-x-Pro)[56–58], while the SH3 domains bind proteins with signature proline-rich sequences (Pro-x-x-Pro-x-Lys/Arg)[59–61]. The signal transduction function of CrkL is attributed to the formation of coordinately protein complexes that bind to the SH2 and the more N-terminal SH3 domain. In particular, the phosphorylation of CrkL on Tyr 207 causes intra-molecular binding of the linker region to the SH2 domain, sequestering the SH2 and N-SH3 and preventing them from binding target proteins[62]. In fact, it's well known that together with the specific binding site recognition, the quaternary structure of a domain has a role in modulating the affinity and allostery during the binding event[63].

1.5 The aim of the thesis

Proteins have evolved increasing their structural complexity by selecting and orchestrating multiple domains[66]. In fact, inter-domain communication is an essential event both for the folding and the binding of these macromolecules. Surprisingly, even though multi-domain proteins compose the majority of the proteome, only recently biophysical studies have been expanded to more complex systems. For this reason, my research will present two examples on how a single domain is influenced by its natural supramodular context, both in the folding and in the binding processes.

The first part of the thesis is focused on the comparison of the folding pathways of the PDZ1-PDZ2 tandem of whirlin and the PDZ1 expressed in isolation. The characterization of the intermediates as well as transition states was performed by using the Φ-values analysis. The analysis of the

mutational data sets together with the *in silico* prediction of the frustration pattern of the PDZ1 domain allowed to pinpoint the residues that act differently in the folding of the two constructs.

The second part of this work aims to describe the differences in the binding features between the CrkL protein and its N-terminal SH3 domain. Thanks to my secondment to professor Ylva Ivarsson's laboratory in the Chemistry Department of Uppsala University, I have screened the CrkL and its constituent domains using the phage display technique. Thirty-nine peptides were identified as possible binders and eventually the interactions between the full-length protein and the isolated N-SH3 with four peptides was validated through equilibrium binding experiment.

The combination of these two studies aims to answer to how domains interact within a full-length protein, and how this interaction can vary the folding and the binding processes of a single domain.

2. Materials and Methods

2.1. Site-directed mutagenesis

2.1.1 PDZ1-PDZ2 tandem and PDZ1 site-directed variants

The constructs encoding the site directed variants of PDZ1 (140-224 residues) and the tandem (140-361 residues) were engineered substituting Y168W (used as pseudo-wild type form). The constructs were obtained inserting the gene encoding pseudowild-type (pWT) in the pET28b+ expression vector. This DNA was chosen as a template to perform site-directed mutagenesis using the QuickChange Lightning Site-Directed Mutagenesis kit (Agilent technologies) according to the manufacturer's instructions. The stability and correct folding of pWT was validated by urea denaturation and circular dichroism experiments. All substitutions were conservative and were confirmed by DNA sequencing.

2.1.2 CrkL and N-SH3 constructs

The sequence encoding for CrkL protein (1-302 residues) and the N-terminal SH3 domain (123-183 residues) were subcloned in pET28b+ expression vector. The constructs were engineered substituting C44S and C248S to avoid any dimerization process. The transformations for DNA extraction were made in DH5α *E.coli* cells and sent for sequencing.

2.2 Protein expression and purification

2.2.1. PDZ1-PDZ2 and PDZ1 mutants

DNA was transformed into *Escherichia coli* BL21 cells for protein expression. Bacterial cells were grown in LB medium, with kanamycin at final concentration of 30 μg/mL. The proteins expression was induced with isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 100

μg/mL at an OD_{600} of 0.8, and the cells were incubated for 3 h at 30 °C. The cells were then cultured after an overnight incubation at 20 °C. Each construct and mutant were purified from the soluble fraction in 50 mM NaPho at pH 7.2 with 0.3 mM NaCl with a HiTrap Chelating High-Performance column (GE Healthcare) and then eluted with a gradient to 1 M imidazole. The imidazole was removed using a HiPrep Desalting column (GE Healthcare). The purity of the proteins sample was confirmed by sodium dodecyl sulfate polyacrylamide gel electrophoresis.

2.2.2. CrkL and N-SH3 constructs

The expression of all the His-tagged constructs was performed in *E. coli* cells, strain BL21. Bacterial cells were grown in LB medium, with 30 μg/mL of kanamycin, at 37 °C until OD_{600} = 0.7−0.8 and then induced with 0.5 mM IPTG. The cultures were grown at 37 °C for three hours after induction, kept at 25 °C overnight and then collected by centrifugation. Purification was performed resuspending the pellet in 50 mM TrisHCl, 0.3 M NaCl, pH 7.5 buffer with the addition of antiprotease tablet (Complete EDTA-free, Roche), and then sonicated and centrifuged. The soluble fraction from bacterial lysate was loaded onto a nickel-charged His-Trap chelating HP (GE Healthcare) column equilibrated with 50 mM TrisHCl, 0.3 M NaCl and pH 7.5. Protein was then eluted with a gradient from 0 to 0.5 M imidazole by using an ÄKTA-prime system. Fractions containing the protein were collected, and the imidazole was removed using a HiTrap Desalting column (GE Healthcare), with the protein purified in the final buffer of TrisHCl 50 mM, NaCl 0.3 M, pH 7.5. The purity of the proteins was analyzed through SDS-page.

2.3 Proteomic peptide-phage display

2.3.1 Construction of the phage libraries

The libraries used in the phage display experiments were already available in prof. Ivarsson laboratory. The second-generation human disorderome (HD2) library encodes peptides representing the intrinsically disordered regions (IDRs) of the intracellular human proteome[67]. These regions were tiled as 16 amino acid long peptides overlapping different section of each protein. HD2 contains 938.427 peptides from 16.969 proteins. The RiboVD library includes peptides of IDRs of proteins from RNA viruses. It contains in total 19.549 unique peptides in 1.074 proteins from 229 strain of 211 viral families[68].

2.3.2 Phage selections and initial NGS data processing

The GST-tagged-proteins (0.5 mg/mL) were coated in 96-well Flat-bottom Immuno Maxisorp plates overnight at 4 °C. In parallel, GST was plated in a preselection plate. The Maxisorp plates were blocked with 0.5% BSA in PBS. The phage library (~ 1012 phage particles in each well) was added to the preselection plate for 1 h, transferred to the target proteins and were allowed to bind for 2 h. Unbound phages were removed by five times washing with cold wash buffer (PBS, 0.5% Tween-20) and bound phages were eluted by direct infection into bacteria by the addition of 100 µL of *E. coli* SS320 in 2YT (OD_{600} = 0.8) to each well and incubation for 30 min at 37 °C with shaking. M13K07 helper phage (NEB, Ipswich, MA, USA) was added to enable phage production, and the cultures were incubated for 45 min at 37 °C with shaking. Eluted phages were amplified overnight in 1.5 mL 2YT supplemented with antibiotics (carbencillin and kanamycin). Bacteria were then pelleted by centrifugation; the supernatant was heat inactivated at 65 °C for 15 min, shilled on ice and then used for the next round of

selections. Five rounds of phage panning were conducted, and the selections were followed by pooled phage enzyme linked immunosorbent assays, which suggested that the selections were saturated after 4 days of selections. Phage pools of round four were barcoded for NGS. Undiluted amplified phage pools (5 μL) were used as templates for 24 cycles, 50 μL PCR reactions using unique combinations of barcoded primers for each reaction and Phusion High Fidelity DNA polymerase (NEB) with a maximum polymerase and primer concentrations. The PCR products were confirmed by gel electrophoresis (2% agarose gel) of 1 μL PCR products. The concentrations of the PCR products were estimated using PicoGreen dye (Invitrogen) and using lambda phage double-stranded DNA (dsDNA; Invitrogen) as a standard. The PicoGreen dye was diluted 1 : 400 in TE buffer and mixed with 1 μL of dsDNA standard or PCR product in a low-fluorescence 96-well plate (Bio-Rad, Hercules, CA, USA). The plate was briefly centrifuged before reading the fluorescence in a qPCR machine (BioRad; excitation 480 nm, emission 520 nm). The blank value was subtracted and the DNA concentration of the sample determined from the standard curve. Equal amounts of each PCR products were pooled. The PCR amplicons (~ 3 μg) was sent to Cofactor Genomics (St. Louis, MO, USA) for NGS.

2.4 Equilibrium binding experiments

Equilibrium experiments were carried out on Fluoromax single-photon counting spectrofluorometer (Jobin-yvon, Newark, NJ, USA) at 10 °C, using a quartz cuvette with a path length of 1 cm, in 50 mM Hepes pH 7.5. The change in fluorescence of the naturally present tryptophanes of the constructs was measured at increasing concentrations of peptides. The excitation wavelength was 280 nm and fluorescence spectra were recorded between 300 and 400 nm. For the experiment with FAM110B peptide (final concentration

of 3 μM) the CrkL and N-SH3 were at fixed concentrations of 1.5 μM; binding experiment with WDR70 peptide (final concentration of 8 μM) was performed with CrkL and N-SH3 at fixed concentration of 1 μM; experiment with HKU5 peptide (final concentration of 40 μM) was performed with CrkL and N-SH3 at fixed concentrations of 1 μM; experiment with NS peptide (final concentration of 3 μM) the proteins had fixed concentrations of 0.5 μM.

2.5. Fluorescence kinetic experiments

Rapid-mixing kinetic folding and unfolding experiments were carried out with SX-18 and PiStar stopped-flow devices (Applied Photophysics). For all the experiments, the excitation wavelength was 280 nm and fluorescence emissions were measured with a 320-nm cut-off filter. Protein final proteins concentration was 1.5 μM. The temperature was set at 25 °C and the buffer used was 50 mM TrisHCl pH 7.5 and 0.3 M Na_2SO_4. Refolding experiments of PDZ1-PDZ2 were performed with the protein diluted in mild denaturant concentration (i.e. 2.2 M GdnHCl).

2.6 Data analysis

2.6.1 Φ-value of PDZ1 and PDZ1-PDZ2 tandem

Kinetic traces were fitted with a single exponential decay using Applied Photophysics software to obtain the observed rate constant k_{obs}. The logarithmic values of k_{obs} were plotted versus GdnHCl concentration. For each experiment, an average calculated from at least 5 independent traces was satisfactorily fitted with a single exponential equation. The chevron plot was fitted using the three-state equation:

$$k_{obs} = \frac{k_{IN}^{0} \exp(-m_{IN}[GdnHCl])}{(1+K_{DI} exp(m_{DI}[GdnHCl])} + k_{NI}^{0} \exp\,(m_{NI}[GdnHCl]) \quad \text{(Eq. 16)}$$

2.6.2 Equilibrium Binding experiment

Data from equilibrium binding experiments were fitted with the following equation:

$$Fluo_{obs} = \left(\left(\frac{Fluo_{max} * [peptide]}{K_D + [peptide]}\right)\right) + cost \quad \text{(Eq. 17)}$$

In which $Fluo_{obs}$ is the observed fluorescence, the $Fluo_{max}$ the asymptot of the function, the K_D the dissociation rate constant of the binding reaction and cost is the minimum value of fluorescence.

3. Results and Discussion

3.1 PART 1: Does PDZ1 fold by the same pathway in isolation and in the PDZ1-PDZ2 tandem?

3.1.1 Φ-value analysis of PDZ1 and PDZ1-PDZ2

The main purpose of this study lies in comparing the folding of a protein domain, PDZ1 from Whirlin, in isolation and in the context of its multi-domain supramodular organization, PDZ1-PDZ2. To achieve this aim, we resorted to perform a complete kinetic analysis of the folding and unfolding properties of these two constructs. In analogy to our previous works[31,32], in all the proteins described in this study, we inserted a mutation, Y168W, whose fluorescence reports efficiently the denaturation of PDZ1, whereas no changes in fluorescence can be observed in the case of PDZ2.

The folding and unfolding kinetics of PDZ1 was investigated by stopped-flow experiments. In all cases, folding and unfolding time courses were fitted satisfactorily to a single exponential decay at any final denaturant concentration. The semi-logarithmic plot of the observed folding/unfolding rate constant versus denaturant concentration is reported in Figure 10.

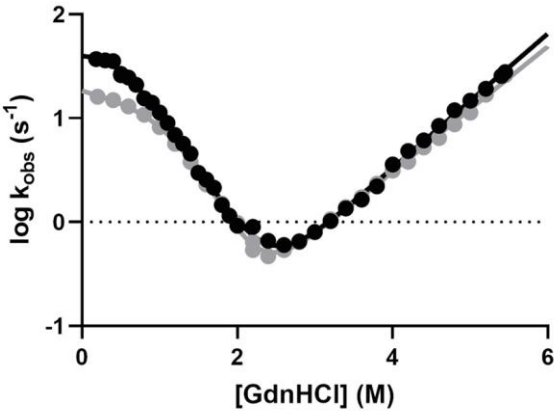


Figure 10. *Semi-logarithmic plot of the observed folding/unfolding rate constant versus denaturant concentration. Chevron plot of PDZ1 (black circles) and PDZ1-PDZ2 (gray circles) both fitted with a three-state equation, which indicates the presence of a folding intermediate. As detailed in the text, the PDZ1-PDZ2 refolding branch was obtained with the protein diluted in mild denaturant concentrations (i.e., 2.2 M GdnHCl) to ensure PDZ2 domain of the tandem to preserve a folded conformation.*

Whereas the logarithm of the unfolding rate constant appears to vary linearly with denaturant concentration, the refolding rate constant displays a deviation from linearity (roll-over effect), which indicates the presence of a folding intermediate. Accordingly, the dependence of the observed rate constant on denaturant concentration was fitted to a three-state equation as formalized in the Experimental Section. The calculated folding and unfolding parameters are reported in Table 1.

Table 1 – *Kinetic and thermodynamic (un)folding parameters of PDZ1 and PDZ1-PDZ2 site-directed variants.*

	k_{IN} (s^{-1})	k_{NI} (s^{-1} 10^{-3})	K_{D-I} (10^{-4})	$\Delta\Delta G_{D-I}$ (kcal mol^{-1})	$\Delta\Delta G_{TS2-N}$ (kcal mol^{-1})	$\Delta\Delta G_{D-N}$ (kcal mol^{-1})	Φ_I	Φ_{TS2}
PDZ1 WT	43±2	15±1	45±10					
PDZ1 L141A	57±5	120±10	400±80	1.30±0.10	1.23±0.03	2.40±0.20	0.55±0.07	0.48±0.04
PDZ1 A147G	56±3	32±1	100±10	0.50±0.10	0.44±0.03	0.80±0.10	0.63±0.16	0.44±0.09
PDZ1 L153A	31±3	47±2	370±70	1.20±0.10	0.66±0.04	2.10±0.20	0.59±0.08	0.69±0.03
PDZ1 I157V	48±3	44±2	80±10	0.30±0.10	0.63±0.04	0.90±0.10	0.36±0.15	0.30±0.12
PDZ1V165A	43±2	20±1	60±10	0.20±0.10	0.15±0.03	0.30±0.10	/	/
PDZ1 I167V	32±1	30±1	80±10	0.30±0.10	0.39±0.03	0.90±0.10	0.35±0.12	0.56±0.07
PDZ1 V169A	5±1	57±4	270±130	1.00±0.30	0.78±0.05	3.10±0.30	0.34±0.10	0.75±0.03
PDZ1 L171A	97±5	13±1	27±3	-0.30±0.10	-0.11±0.04	-0.90±0.10	0.34±0.13	0.87±0.04
PDZ1 L177A	38±2	20±1	75±10	0.30±0.10	0.16±0.03	0.50±0.10	0.56±0.23	0.70±0.09
PDZ1 A178G	15±1	90±2	290±30	1.10±0.10	1.04±0.03	2.80±0.10	0.40±0.04	0.62±0.02
PDZ1 V185A	34±2	38±2	100±20	0.50±0.10	0.54±0.04	1.20±0.10	0.42±0.12	0.54±0.06
PDZ1 I189V	28±1	50±1	100±10	0.50±0.10	0.70±0.03	1.40±0.10	0.33±0.08	0.51±0.04
PDZ1 L190A	33±2	74±2	230±40	1.00±0.10	0.93±0.03	2.10±0.10	0.47±0.07	0.55±0.03
PDZ1 V192A	29±1	34±1	200±20	0.90±0.10	0.47±0.03	1.60±0.10	0.55±0.08	0.70±0.03
PDZ1 L197A	34±2	290±10	270±60	1.10±0.10	1.74±0.03	2.90±0.20	0.36±0.06	0.41±0.03
PDZ1 A198G	47±2	43±1	80±10	0.30±0.10	0.61±0.03	0.90±0.10	0.37±0.13	0.32±0.09
PDZ1 A203G	38±3	36±2	120±20	0.60±0.10	0.51±0.04	1.20±0.20	0.50±0.13	0.57±0.06
PDZ1 A205G	26±5	140±10	400±210	1.30±0.30	1.31±0.05	2.90±0.40	0.44±0.12	0.55±0.06
PDZ1 V206A	39±5	42±3	260±70	1.00±0.20	0.59±0.04	1.70±0.20	0.61±0.13	0.65±0.05
PDZ1 A208G	28±7	280±20	1100±600	1.90±0.30	1.71±0.05	3.80±0.40	0.49±0.10	0.56±0.05
PDZ1 L215A	26±4	68±5	260±100	1.00±0.20	0.88±0.05	2.20±0.30	0.47±0.12	0.60±0.05
PDZ1 V216A	28±3	57±3	200±40	0.90±0.10	0.78±0.04	1.80±0.20	0.51±0.09	0.56±0.05

	k_{IN} (s^{-1})	k_{NI} (s^{-1} 10^{-3})	K_{D-I} (10^{-4})	$\Delta\Delta G_{D-I}$ (kcal mol^{-1})	$\Delta\Delta G_{TS2-N}$ (kcal mol^{-1})	$\Delta\Delta G_{D-N}$ (kcal mol^{-1})	Φ_I	Φ_{TS2}
P1P2 WT	23±1	13±1	16±1					
P1P2 L141A	23±2	63±4	20±10	0.20±0.30	0.92±0.04	1.20±0.30	0.20±0.22	0.21±0.18
P1P2 A147G	16±1	18±1	10±2	-0.30±0.10	0.19±0.03	0.10±0.10	/	/
P1P2 L153A	9±1	31±2	20±10	0.10±0.30	0.50±0.04	1.10±0.30	/	/

P1P2 I157V	11±1	23±2	10±4	-0.30±0.30	0.32±0.05	0.50±0.30	-0.68±0.76	0.29±0.47
P1P1 V165A	17±1	16±1	10±2	-0.20±0.10	0.13±0.03	0.10±0.10	/	/
P1P2 I167V	16±1	21±1	10±2	-0.20±0.10	0.29±0.03	0.30±0.20	/	/
P1P2 V169A	2±1	59±3	130±70	1.30±0.30	0.89±0.03	3.50±0.30	0.35±0.09	0.75±0.02
P1P2 L171A	58±5	10±1	10±3	-0.30±0.20	-0.17±0.05	-1.00±0.20	0.28±0.20	0.83±0.06
P1P2 L177A	15±1	12±1	40±10	0.60±0.10	-0.03±0.03	0.80±0.20	0.74±0.22	1.04±0.04
P1P2 A178G	7±1	100±5	400±160	1.90±0.30	1.21±0.03	3.80±0.30	0.50±0.07	0.68±0.02
P1P2 V185A	13±1	30±1	50±10	0.70±0.20	0.48±0.03	1.50±0.20	0.47±0.11	0.69±0.04
P1P2 I189V	11±1	35±3	10±10	-0.20±0.30	0.58±0.04	0.90±0.30	-0.20±0.40	0.30±0.30
P1P2 L190A	15±1	58±2	80±10	1.00±0.10	0.88±0.02	2.10±0.10	0.47±0.06	0.58±0.03
P1P2 V192A	11±1	26±1	70±10	0.90±0.10	0.40±0.03	1.70±0.10	0.51±0.07	0.77±0.02
P1P2 L197A	16±2	270±10	240±100	1.60±0.30	1.77±0.03	3.60±0.30	0.44±0.08	0.51±0.04
P1P2 A198G	22±2	47±2	30±10	0.50±0.20	0.75±0.03	1.20±0.20	0.37±0.16	0.40±0.10
P1P2 A203G	15±1	32±2	40±10	0.60±0.20	0.52±0.03	1.30±0.20	0.42±0.13	0.61±0.05
P1P2 A205G	9±1	86±6	130±60	1.30±0.30	1.11±0.04	2.90±0.30	0.43±0.11	0.62±0.04
P1P2 V206A	14±1	36±2	50±10	0.60±0.20	0.60±0.03	1.50±0.20	0.41±0.13	0.61±0.05
P1P2 A208G	5±1	150±20	50±80	0.70±0.90	1.45±0.07	3.00±0.90	0.24±0.32	0.52±0.15
P1P2 L215A	13±1	39±3	20±10	0.10±0.30	0.64±0.04	1.10±0.30	0.13±0.24	0.44±0.14
P1P2 V216A	18±1	54±3	70±20	0.80±0.20	0.84±0.03	1.80±0.20	0.46±0.09	0.54±0.04

We previously showed that the equilibrium unfolding of PDZ1-PDZ2 proceeds in a stepwise manner and the first PDZ domain may be denatured at 'mild denaturant concentration', i.e. 2.2 M of GdnHCl, whereas the second PDZ domain retains its native conformation[31,32]. Thus, in an effort to compare the folding of PDZ1, in isolation and in the context of its multi-domain supramodular organization, we measured the chevron plot of PDZ1-PDZ2 by first denaturing the protein in 2.2 M GdnHCl and then triggering refolding by rapidly mixing with buffer in the presence of increasing concentrations of GdnHCl. Conversely, unfolding was measured by challenging native PDZ1-PDZ2 with GdnHCl at different concentrations. Also in this case, both folding and unfolding time courses were satisfactorily fitted to a single exponential decay at any final denaturant concentration. A superposition between the chevron plots of PDZ1 in isolation and in the

tandem is reported in Fig. 10. Whilst the unfolding of the two proteins appears essentially identical, it might be noticed that inclusion of PDZ1 in a tandem repeat result in a minor but detectable stabilization of the folding intermediate, as probed by the slightly more pronounced roll-over effect in the refolding branch. Importantly, this intermediate is distinct from the major misfolding events previously reported that may occur only when both PDZ domains are unfolded in PDZ1-PDZ2 and cannot be detected in PDZ1 in isolation. In analogy to what reported above for PDZ1, data were fitted to a three-state folding equation and the resulting parameters are listed in Table 1.

To address the details of the folding of PDZ1 we performed Φ-value analysis. By following this methodology, residue-specific structural information of metastable state(s) along the reaction pathway is inferred by comparing the kinetics of folding of the wild-type protein with those of a series of conservative single mutants[44,69]. Quantitatively, the strength of the contacts is measured by the Φ-value, which normalizes the stability change of the metastable state upon mutation to that of the native state. A Φ-value close to 1 is indicative of native-like structure in the metastable state of that specific residue, whereas a Φ value equal to 0 suggests that the mutated residue is as unstructured in the metastable state as it is in the denatured state.

Twenty-eight site-directed variants of PDZ1 were produced, expressed and subjected to (un)folding experiments. The mutants were designed according to standard rules of Φ-value analysis, which were extensively discussed elsewhere[44,69–71]. In summary, a conservative deletion of hydrophobic side chains was designed, a type of mutation that represents the easiest to be interpreted. The chevron plot obtained for each variant is reported in Figure 11.

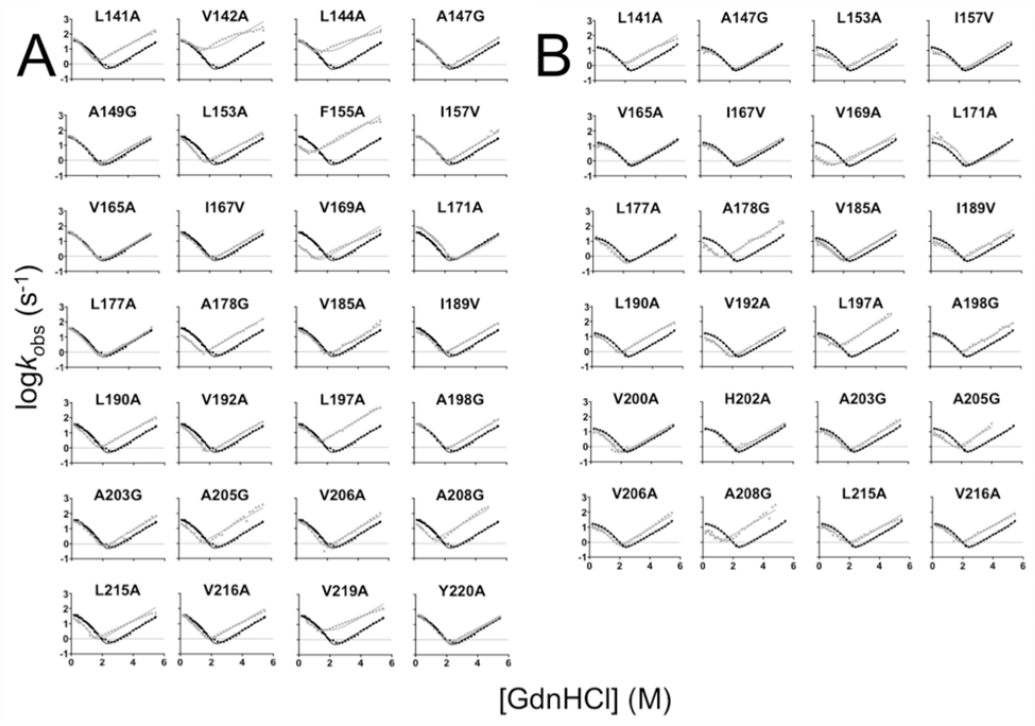


Figure 11. *Chevron plots of the 52 site-directed mutants. The PDZ1 wt (Panel A) and the PDZ1-PDZ2 wt (Panel B) are reported in black and their mutated variants in gray. Data were fitted with equation 16, as formalized in Material and Methods section. The buffer used in all the experiments was 50 mM TrisHCl pH 7.5 and 0.3 M Na2SO4 and the temperature was 25°C.*

In analogy to previous works on multistate systems[40,72,73], the obtained chevron plots were globally fitted to a three-state equation with shared m-values.

To provide structural information of the intermediate and transition states, the mutants were divided in three groups based on their measured Φ-values: small (Φ < 0.3; red), intermediate (0.3 < Φ < 0.7; magenta), and large (Φ > 0.7; blue). The color-coded mutations were then mapped into the structure of PDZ1 (Fig. 12).

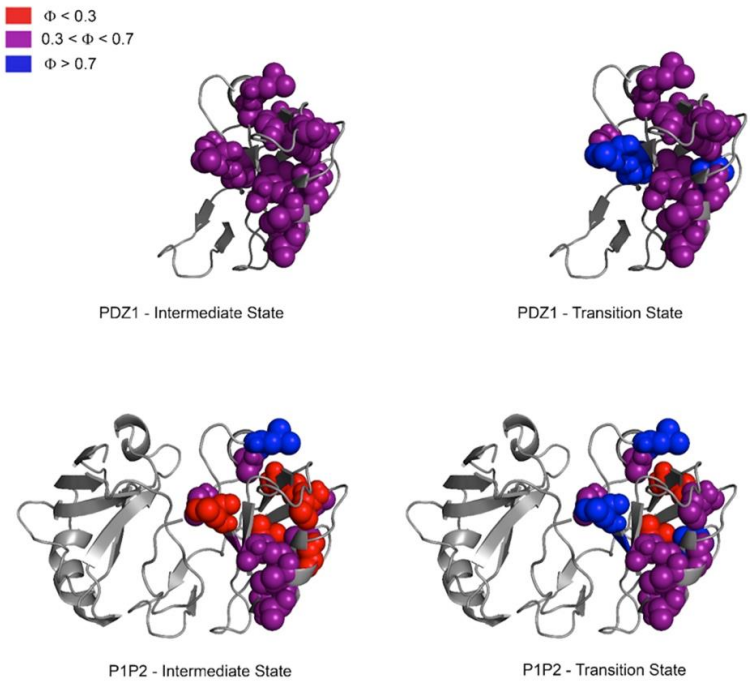


Figure 12. *Φ-values mapped on the PDZ1 and PDZ1-PDZ2 3D structures. Color-coded distribution of Φ-values is calculated for the intermediate state and transition state on the PDZ1 and PDZ1-PDZ2 structures (Φ < 0.3; red) 0.3 < Φ < 0.7; magenta Φ > 0.7; blue). While the intermediate states appear to be more malleable and characterized by a lower degree of native-like structure in PDZ1-PDZ2 compared to PDZ1, this difference is less pronounced in the comparison of transition states.*

The structural distribution of the measured values for the PDZ1 suggests the intermediate to be characterized by the formation of a weak nucleus, which appears to be diffused in the whole globule. Conversely, native structure in the PDZ1 transition state presents a more polarized distribution, encompassing mostly the residues found in the β3-β4 strands.

To test the robustness of the folding of PDZ1 in the context of the muti-domain supramodular structure we resorted to perform a comparative Φ-value analysis. Hence, in analogy to what described above, we expressed, purified and characterized twenty-four site directed mutants and subjected them to kinetic folding and unfolding experiments.

Figure 11 reports a mutant-by-mutant comparison of the chevron plots measured for each of the PDZ1 in isolation and in the PDZ1-PDZ2 tandem, with the calculated folding and unfolding parameters listed in Table 1. Overall, it might be observed that in essentially all cases, the chevron plots of all the variants appear nearly identical in both constructs, with the relevant exception of the refolding roll over that appears more pronounced in the case of the PDZ1-PDZ2 constructs rather than in the case of PDZ1 in isolation.

In the PDZ1-PDZ2 construct, the architecture of the intermediate is less structured compared to the isolated domain, as shown by the presence of lower Φ-values. This finding contrasts what observed for the transition state, which is rather robust and structurally similar to what observed in the case of PDZ1 in isolation. Thus, while in the late stages the two contructs display the same robust pattern, in the early events the mutational analysis reveals significant differences between PDZ1 in isolation and in tandem.

3.1.2 Frustration pattern of PDZ1

A direct way to compare mutational data sets is to perform Φ-Φ plots of a relevant state, as well as in comparing the changes in free energies upon mutation[70]. Figure 13 depicts the Φ-Φ and ΔΔG plots for the intermediate and transitions states of PDZ1 in isolation and in the PDZ1-PDZ2 construct.

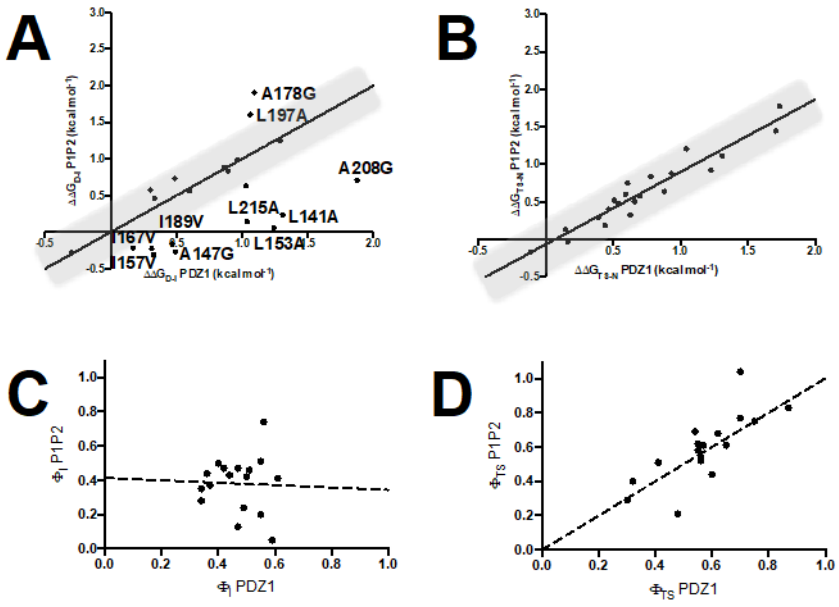


Figure 13. *ΔΔG plots and Φ-Φ plots for PDZ1 and PDZ1-PDZ2 intermediate states (Panels A and C) and transitions states (Panels B and D). Each point in the graphs represents a single site-directed mutation occurring in both proteins. While for the intermediate states there is a pronounced scattered distribution, with several variants lying outside of the linear fit, a strong linear correlation is evident for the analysis of the transition states.*

It is evident that, whilst the data for the transition states are conserved in the two constructs, consistent with a linear correlation and a slope of 1, in the case of the intermediate there are clear differences between the two

constructs. Thus, there is an intriguing picture emerging from the comparison between PDZ1 in isolation and in the context of its supramodular organization that suggests a clear robustness to characterize the late stages of folding, whereas a more malleable behavior may be detected in the early stages.

Thus, whilst the late events appear strongly committed to the native topology, the early events are more malleable and prone to changes depending on the presence/absence of the adjacent domain. We note that such behavior parallels what expected from the funneled energy landscape theory that postulates the native bias to be weak at early stages of folding, allowing for alternative early folding events.

Ferreiro, Wolynes and co-workers provided a public algorithm that allows calculating the frustration patterns of proteins (available at http://frustratometer.qb.fcen.uba.ar)[74]. On the light of what summarized above, to explain the observed differences in the folding intermediate of PDZ1 in isolation and in the context of its multidomain structure, we calculated the frustration pattern of PDZ1. Figure 14 depicts the structure of PDZ1 and highlights the frustration patterns within this protein domain along with the residues that are prone to structural changes in the folding intermediate. Strikingly, we found a remarkable superposition between residues prone to alternative folding pathways and the frustrated regions of PDZ1. This finding further confirms that frustration sculpts the early stages of folding, whereas it has little effects on the late stages of the reaction.

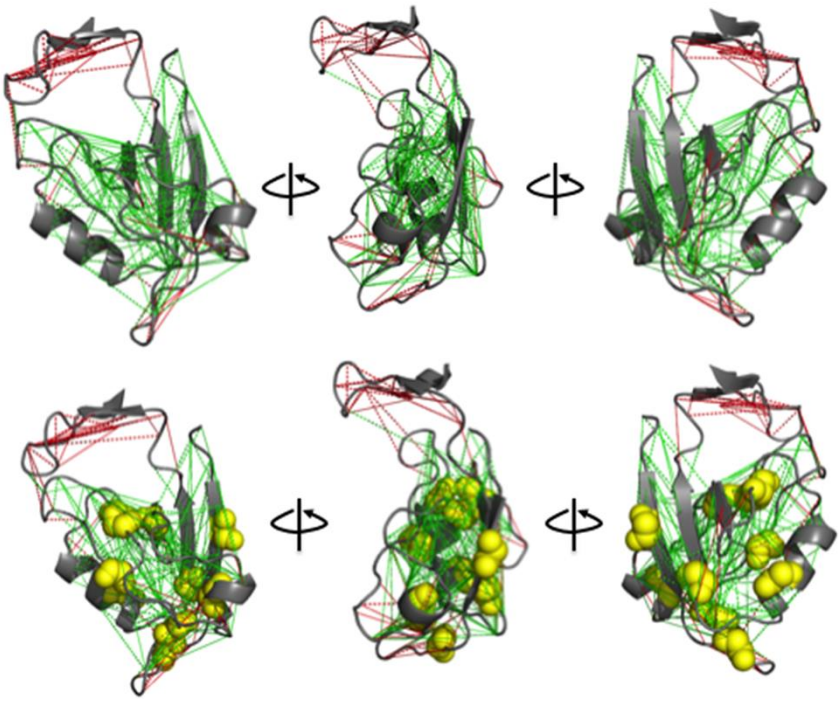


Figure 14. *Frustration pattern of PDZ1 was calculated from the algorithm kindly provided by Wolynes, Ferreiro and co-workers and available at the website http://frustratometer.qb.fcen.uba.ar. The red lines indicate local frustrated patterns within the structure. The yellow spheres are the residues that show a detectable difference in* $\Delta\Delta G_{D\text{-}I}$ *highlighted in Figure 13 panel A.*

3.2 PART 2: Are there any differences in the binding affinities between the N-SH3 domain and the full-length CrkL?

3.2.1 Phage-display selections

CrkL and its constituent domains (SH2, N-SH3, C-SH3 and the NSH3CSH3 tandem) were used as bait proteins against two libraries that display peptides from the IDRs of the intracellular human proteome (HD2) and peptides from IDRs of proteins of RNA viruses (RiboVD). The selections were confirmed successful by pooled phage ELISA as shown by Figure 15. The measured absorbance at 450 nm displayed an increase of the values from the first to the fourth round of amplification, indicating the presence of enriched binding phages. In particular, the full-length protein, the NSH3CSH3 tandem and the N-terminal SH3 domain showed the higher binding activity compared to the C-terminal SH3 domain. Whereas the SH2 domain escaped from the PD analysis, suggesting that this domain may not recognize IDP as ligands. In fact, even though a recent work on the SH2 interactome has highlighted unusual binding characteristics[75], the presence of the pTyr seems to be necessary for SH2 cellular interactions[56].

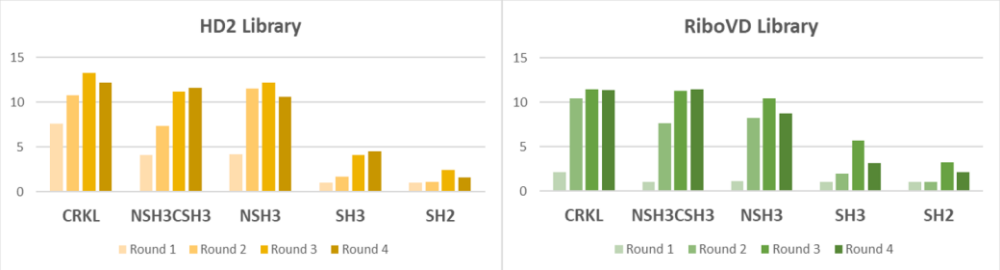


Figure 15. *Phage pool ELISA normalized absorbance (450 nm) for HD2 and the RiboVD libraries. The phage display libraries were screened through four cycles of enrichment and amplification. At each cycle the phages were*

tested for their ability to bind to the bait proteins using an ELISA assay as described in Materials and Methods section.

The fourth round of selection was sequenced with NGS, and the resulted pool of peptide ligands were filed in an online toolkit available at http://slim.icr.ac.uk/proppd/. The software can analyse the selected peptides and therefore identify the consensus motif for each bait protein. This post-process of the PD data allows the identification of short linear motif (SLiM) interactions. The selected peptides for all the constructs (except for the SH2) were dominated by sequences containing the shared consensus motif x-P-P-L/V-P-P-K/R (Figure 16). The identification of preferred interaction of the SH3 domain by phage-display was already described in the literature[76–78] in which is reported a more general binding motif R-P-L-P-x-x-P. Our results on the CrkL protein revealed a more specific binding consensus, involving the lysine in position +2 and a hydrophobic residue in position -1. In particular, the lysine has been demonstrated to be of key importance for binding specificity of the CrkL N-SH3 because of its ionic interaction with three negatively charged residues in the binding pocket of the domain[79].

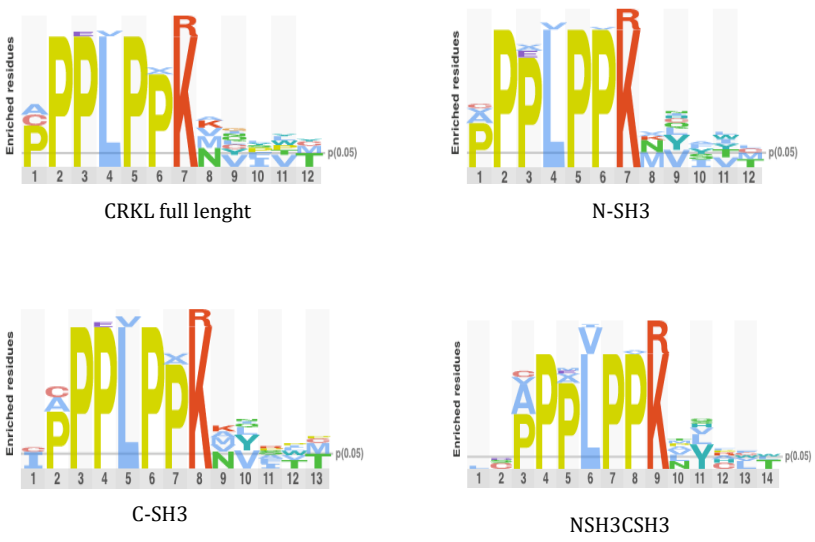


Figure 16. *Consensus motif identified for CrkL protein, the N-terminal and the C-terminal SH3 domain and the NSH3CSH3 tandem. The NGS converged in a common consensus of P-P-L-P-P-K for all the constructs. The letter dimension is proportional to the residue's enrichment.*

The results of the selection identified around 30 enriched peptides for CrkL full-length, 24 from the HD2 library and 6 for the RiboVD. Same results were obtained for the N-SH3 domain and the tandem. Whereas the C-SH3 data were of relatively poor quality, with just two peptides occurring in replicate selections, suggesting that the C-terminal domain of CrkL probably is not essential in the interaction with the intracellular IDPs. This finding is corroborated by Posern and coworker's work in which the binding capacity of the CrkL C-terminal domain was tested with no results achieved[80].

Eventually, we investigated the presence of known binders for the CrkL constructs by comparing the data set with ligands reported in BioGrid database[81]. Strikingly we found that only 50% of identified peptides were

from proteins previously reported to interact with our system, meaning that half of the selected ligands are yet to be characterized. All the interactors are listed in Table 2 and in the comment section are reported the interactions already studied in literature. To assess the binding quantitatively, we tested 4 unknown peptides, selected according to their physiological relevance, the specificity and the percentage of occurrence in the NGS.

Considering the CrkL implication in the amplification of several cancer cell lines, we selected peptides, from the result of the HD2 library screening, which are part of proteins that are down/up regulated in cancer cell (FAM110B protein[82] and WD repeat-containing protein 70[83]). Instead regarding the RiboVD peptides we selected influenza A virus non-structural protein 1 (NS) and the nucleoprotein Bat-coronavirus (HKU5).

Table 2 – *CrkL binding sequences derived from screening the HD2 (orange table) and RiboVD (green table) phage display libraries. All the phage-derived sequences are listed with the associated protein. In the comment section is reported the reference of the corresponding interaction.*

Hit	Protein	Comments/references
HEESDAPLLPPRVYST	Adhesion G protein-coupled receptor L3 ADGRL3	
DAPLLPPRVYSTENHQ		
DPVLPPLPAKRHLAEL	Arf-GAP with Rho-GAP domain, ANK repeat and PH domain-containing protein 1 ASAP2	Huttin EL (2021)
PPLPAKRHLAELSVPP		
PPLPPRNVGKVQTASS	Arf-GAP with SH3 domain, ANK repeat and PH domain-containing protein 2	Huttin EL et al. (2021)
DCPPPLPVKNSSRTLV	Arf-GAP with SH3 domain, ANK repeat and PH domain-containing protein 3	Huttin EL et al. (2021)
DVADVPPPLPLKGSVA	Dedicator of cytokinesis protein 1 DOCK1	Hein et al. (2015)
MTGADVADVPPPLPLK		
EPPALPPRTLEGLQVE	Hematopoietic lineage cell-specific protein HCLS1	
NLPPPLPPKKYAITSV	Inactive ubiquitin carboxyl-terminal hydrolase 53	Luck et al. (2017)

	USP53	
PPPPPPPLPEKKLITR	Leiomodin-2 LMOD2	
VGGQDGEQAPPLPIKA	Low-density lipoprotein receptor-related protein 10 LRP10	
IAPPVPLKAQTVTDSM	Mitotic checkpoint serine/threonine-protein kinase BUB1	
TPHHPPALPSKLPTEV	Msx2-interacting protein	
PPALPSKLPTEVNHVP	SPEN	
HPELPTKGKDVSYCPV	NEDD4-binding protein 2 N4BP2	
PPALPPKPPKPMTSAV	Phosphatidylinositol 3-kinase regulatory subunit gamma PIK3R3	Mueller PM et al. (2020)
DTVESVVSPPELPPRN	Phosphatidylinositol 3,4,5-trisphosphate 5-phosphatase 1 INPP5D	Arai A et al. (2001)
SVVSPPELPPRNIPLT		
PPELPPRNIPLTASSC		
PPPLPPRASIWDTPPL	Phosphatidylinositol 4-phosphate 3-kinase C2 domain-containing subunit beta PIK3C2B	Huttlin EL et al. (2021)
PHSSPPPLPAKASRQL	Pleckstrin homology-like domain family B member 1 PHLDB1	family A Thalappilly S et al. (2008)
PPPLPAKASRQLQVYR		
SAPPLPPKPKIAAIAS	Protein FAM110B	
ELPPKKRYYCELCLDD	Ras/Rap GTPase-activating protein SynGAP	known for Crk Birge R et al. (2009)
PPPLPPKNVPATPPRT	SH3-containing GRB2-like protein 3-interacting protein 1 SGIP1	known with Grb2 Park RK (1999)
PTGTPPPLPPKNVPAT		
APVLPGKTGPTVTQVK	Treacle protein TCOF1	
PGLQCPPLPPRVGLPT	Vacuolar protein sorting-associated protein 37B VPS37B	

SETPPPLPPKSPSFQA	WAS/WASL-interacting protein family member 3 WIPF3	Oda A et al. (2001)
PPLPPKMVGKPVNFME	WD repeat-containing protein 70 WDR70	
SSDDELIGPPLPPKMV		
ELIGPPLPPKMVGKPV		

Hit	Protein	Comments/references
VTGLLWLCCLFTPLSM	Replicase polyprotein 1ab Equine virus	
PATSQMEDVPELPPKQ	Nucleoprotein Bat-coronavirus HKU5	
ASLPKLPGKFLQYTVG	Nucleoprotein Bornavirus	
LPDMIQDTPPPVPRKN	Nsp2 Replicase polyprotein 1ab Porcine syndrome virus	
ENGGPSLPPKQKRYMA	Non-structural protein 1 Influenza A virus	Heikken et al. (2008)
PIPLPPKVLENGPNAW	Replicase polyprotein 1ab Porcine virus	

3.2.2 Validating the interactions - Equilibrium binding experiment

The peptide ligands obtained from the phage selection of the full-length protein appeared to interact with both the N-SH3 and the NSH3-CSH3 tandem. This result together with the poor enrichment of selected phages for the C-SH3 and the SH2 support the theory that probably only the CrkL N-terminal domain is the main responsible for the interaction with IDRs of the intracellular proteome. Thus, the next step was to validate the set of interactions resulted from the phage selection, quantifying the differences in the binding affinities between the single domain and the full-length protein.

We determined *in vitro* affinities performing fluorescence equilibrium experiments, measuring the change in fluorescence of the naturally present tryptophanes of the constructs at increasing concentrations of peptides.
The affinities (Table 3) were in the low micromolar range (0.2–14 μM), which is typical for SH3 mediated interactions[80] and similar to what have been observed for synthetic ligands derived from combinatorial phage libraries[76–78].

Table 3 – *Dissociation constants of the CrkL and the N-SH3 domain with selected peptides as determined by fluorescence measurements. Peptide motif indicates just six out of the fifteen residues forming the whole peptides.*

								K_D (μM)	
Protein	**Peptide motif**							**CRKL**	**N-SH3**
Human	-3	-2	-1	0	1	2	3		
FAM110B	**P**	P	L	**P**	P	**K**	P	0.30 ± 0.05	2.00 ± 0.30
WDR70	**P**	P	L	**P**	P	**K**	M	2.40 ± 0.70	1.00 ± 0.25
Viral									
NS	**P**	S	L	**P**	P	**K**	Q	0.20 ± 0.02	0.35 ± 0.10
HKU5	**P**	E	L	**P**	P	**K**	Q	3.50 ± 0.70	13.60 ± 4.40

All the titration curves reported in Figure 17 were fitted satisfactorily equation 17. The experimental validation of the binding activity has shown an expected trend in the dissociation constants between the two constructs, in particular, the binding of CrkL protein with the peptides ligands presents a lower K_D respect to the isolated domain. The higher affinity of the full-length protein respect to the N-SH3 support the idea that these ligands are relevant in a cellular system in the interaction with the full-length protein. This trend is general for all the tested peptide except for WDR70, which seems to have an opposite result.

FAM110B, WDR70 and HKU5 may be novel CrkL binders biologically

relevant for the cellular signalling network. Instead, the interaction with NS protein was already determined as 0.41± 0.05 μM using Biolayer interferometry[84], the same order of magnitude to what observed in the present study.

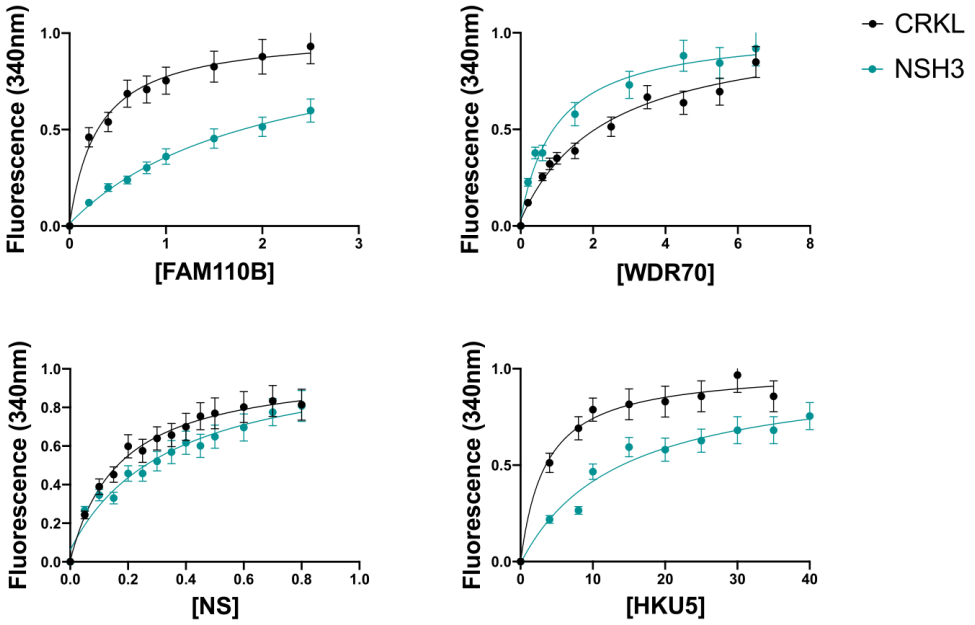


Figure 17. *Equilibrium binding experiment of CrkL full-length (black lines) and N-SH3 domain (green lines) against the four peptide ligands.*

4. Conclusions

Traditionally, much of the theoretical and experimental work in the biophysical characterization of proteins have been focused on globular and isolated domains. The reason of such downsizing is not due to the challenging interpretations of the experimental data, but also to the assumption that the folding and the binding feature of a single domain can be summed up to describe a more general system[85]. For this reason, the main objective of the present work is to highlight the differences in the folding and the binding processes between a multidomain construct and its constituent domains expressed in isolation.

The comparison of the folding of the PDZ1 domain of whirlin with the PDZ1-PDZ2 tandem allowed to measure the differences in the folding pathway of the two constructs. In particular, thanks to the Φ-value analysis we characterized the species involved in the folding, such as intermediates and transition state, giving an additional glimpse in the folding mechanism of this protein. In fact, although the late stages of the folding process are committed to the native state, thus no differences were observed between the two constructs, in the early events the presence of the second PDZ domain seems to increase the frustration of the system, opening the structure to the possibility of engaging misfolding states.

During the secondment in the Chemistry Department of Uppsala university under the supervision of prof. Ylva Ivarsson I had the opportunity to screen the three-domain protein CrkL with the proteomic peptide-phage display technique. This high-throughput method allowed the finding of motif-based interactions that are crucial for signaling and other cellular processes but are difficult to detect because of their low to moderate affinities[86]. We identified known and unknown binders for the full-length protein and for the single

domains. A first semi-quantitative difference was obtained by the NGS results, showing that the ability to bind physiological ligand is performed only by the constructs displaying the N-terminal SH3 domain.

Moreover, we validate the novel interactions between our system and four ligands. The measured dissociation constants were in the low micromolar range (0.2–14 μM), indicating a strong interaction between the CrkL and the selected peptides. Finally, we demonstrated that the presence of tethered domains increases the ability to bind the partners for all the peptides tested excepting for WDR70.

Future work will be aimed to study the folding of wider multidomain constructs, with the focus in pinpointing the interactions that are responsible for the occurrence of misfolded species. In parallel, in the protein-protein interaction field, our work will continue in the characterization of the inter-domain interactions that may be crucial for the binding process.

Eventually, a more general description of proteins may not be achieved without merging the information obtained from the folding and the binding, or rather from the frustration pattern and the allosteric network of a biological system.

5. References

1. Dill, K. A. & MacCallum, J. L. The protein-folding problem, 50 years on. *Science* **338**, 1042–1046 (2012).
2. Ezkurdia, I. *et al.* Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human Molecular Genetics* **23**, 5866–5878 (2014).
3. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973).
4. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
5. Tanford, C. Protein denaturation. *Adv Protein Chem* **23**, 121–282 (1968).
6. Adesnik, M. & Levinthal, C. Synthesis and maturation of ribosomal RNA in Escherichia coli. *J Mol Biol* **46**, 281–303 (1969).
7. Wolynes, P. G. Recent successes of the energy landscape theory of protein folding and function. *Q Rev Biophys* **38**, 405–410 (2005).
8. Yan, Z. & Wang, J. Funneled energy landscape unifies principles of protein binding and evolution. *Proc Natl Acad Sci U S A* **117**, 27218–27223 (2020).

9. Wolynes, P. G., Eaton, W. A. & Fersht, A. R. Chemical physics of protein folding. *Proc Natl Acad Sci U S A* **109**, 17770–17771 (2012).
10. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. Theory of protein folding: the energy landscape perspective. *Annu Rev Phys Chem* **48**, 545–600 (1997).
11. Ferreiro, D. U., Komives, E. A. & Wolynes, P. G. Frustration, function and folding. *Curr Opin Struct Biol* **48**, 68–73 (2018).
12. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**, 536–540 (1995).
13. Apic, G., Gough, J. & Teichmann, S. A. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *Journal of Molecular Biology* **310**, 311–325 (2001).
14. Ekman, D., Björklund, Å. K., Frey-Skött, J. & Elofsson, A. Multi-domain Proteins in the Three Kingdoms of Life: Orphan Domains and Other Unassigned Regions. *Journal of Molecular Biology* **348**, 231–243 (2005).
15. Gerstein, M. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Folding and Design* **3**, 497–512 (1998).

16. Liu, J. & Rost, B. Sequence-based prediction of protein domains. *Nucleic Acids Res* **32**, 3522–3530 (2004).

17. Teichmann, S. A., Chothia, C. & Gerstein, M. Advances in structural genomics. *Current Opinion in Structural Biology* **9**, 390–399 (1999).

18. Batey, S. & Clarke, J. Apparent cooperativity in the folding of multidomain proteins depends on the relative rates of folding of the constituent domains. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 18113–18118 (2006).

19. Batey, S., Scott, K. A. & Clarke, J. Complex Folding Kinetics of a Multidomain Protein. *Biophysical Journal* **90**, 2120–2130 (2006).

20. Borgia, A. *et al.* Transient misfolding dominates multidomain protein folding. *Nat Commun* **6**, 8861 (2015).

21. Borgia, M. B. *et al.* Single-molecule fluorescence reveals sequence-specific misfolding in multidomain proteins. *Nature* **474**, 662–665 (2011).

22. Han, J.-H., Batey, S., Nickson, A. A., Teichmann, S. A. & Clarke, J. The folding and evolution of multidomain proteins. *Nat Rev Mol Cell Biol* **8**, 319–330 (2007).

23. Kumar, V. & Chaudhuri, T. K. Spontaneous refolding of the large multidomain protein malate synthase G proceeds through misfolding traps. *Journal of Biological Chemistry* **293**, 13270–13283 (2018).

24. Tian, P. & Best, R. B. Structural Determinants of Misfolding in Multidomain Proteins. *PLoS Comput Biol* **12**, e1004933 (2016).

25. Arora, P., Hammes, G. G. & Oas, T. G. Folding mechanism of a multiple independently-folding domain protein: double B domain of protein A. *Biochemistry* **45**, 12312–12324 (2006).

26. Pagano, L. *et al.* Probing the Effects of Local Frustration in the Folding of a Multidomain Protein. *Journal of Molecular Biology* **433**, 167087 (2021).

27. Jaenicke, R. Stability and stabilization of globular proteins in solution. *J Biotechnol* **79**, 193–203 (2000).

28. Batey, S., Nickson, A. A. & Clarke, J. Studying the folding of multidomain proteins. *HFSP Journal* **2**, 365–377 (2008).

29. Batey, S. & Clarke, J. The folding pathway of a single domain in a multidomain protein is not affected by its neighbouring domain. *J Mol Biol* **378**, 297–301 (2008).

30. Sánchez, I. E., Morillas, M., Zobeley, E., Kiefhaber, T. & Glockshuber, R. Fast folding of the two-domain semliki forest virus

capsid protein explains co-translational proteolytic activity. *J Mol Biol* **338**, 159–167 (2004).

31. Gautier, C. *et al.* Hidden kinetic traps in multidomain folding highlight the presence of a misfolded but functionally competent intermediate. *Proc Natl Acad Sci U S A* **117**, 19963–19969 (2020).

32. Visconti, L. *et al.* Folding and Misfolding of a PDZ Tandem Repeat. *Journal of Molecular Biology* **433**, 166862 (2021).

33. Lafita, A., Tian, P., Best, R. B. & Bateman, A. Tandem domain swapping: determinants of multidomain protein misfolding. *Current Opinion in Structural Biology* **58**, 97–104 (2019).

34. Petersen, M. & Barrick, D. Analysis of Tandem Repeat Protein Folding Using Nearest-Neighbor Models. *Annu. Rev. Biophys.* **50**, 245–265 (2021).

35. Kloss, E. & Barrick, D. Thermodynamics, Kinetics, and Salt dependence of Folding of YopM, a Large Leucine-rich Repeat Protein. *Journal of Molecular Biology* **383**, 1195–1209 (2008).

36. Kloss, E., Courtemanche, N. & Barrick, D. Repeat-protein folding: new insights into origins of cooperativity, stability, and topology. *Arch Biochem Biophys* **469**, 83–99 (2008).

37. Jackson, S. E. & Fersht, A. R. Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. *Biochemistry* **30**, 10428–10435 (1991).

38. Kelly, C. & Gage, M. J. Protein Unfolding: Denaturant vs. Force. *Biomedicines* **9**, 1395 (2021).

39. Myers, J. K., Pace, C. N. & Scholtz, J. M. Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci* **4**, 2138–2148 (1995).

40. Parker, M. J., Spencer, J. & Clarke, A. R. An Integrated Kinetic Analysis of Intermediates and Transition States in Protein Folding Reactions. *Journal of Molecular Biology* **253**, 771–786 (1995).

41. Baldwin, R. L. On-pathway versus off-pathway folding intermediates. *Fold Des* **1**, R1-8 (1996).

42. Roder, H. & Colón, W. Kinetic role of early intermediates in protein folding. *Curr Opin Struct Biol* **7**, 15–28 (1997).

43. Eyring, H. The Activated Complex in Chemical Reactions. *The Journal of Chemical Physics* **3**, 107–115 (1935).

44. Fersht, A. R., Matouschek, A. & Serrano, L. The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J Mol Biol* **224**, 771–782 (1992).

45. Phizicky, E. M. & Fields, S. Protein-protein interactions: methods for detection and analysis. *Microbiol Rev* **59**, 94–123 (1995).

46. Davey, N. E., Cyert, M. S. & Moses, A. M. Short linear motifs - ex nihilo evolution of protein regulation. *Cell Commun Signal* **13**, 43 (2015).

47. Smith, G. P. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* **228**, 1315–1317 (1985).

48. Smith, G. P. Phage Display: Simple Evolution in a Petri Dish. 22.

49. Behjati, S. & Tarpey, P. S. What is next generation sequencing? *Arch Dis Child Educ Pract Ed* **98**, 236–238 (2013).

50. Mburu, P. *et al.* Defects in whirlin, a PDZ domain molecule involved in stereocilia elongation, cause deafness in the whirler mouse and families with DFNB31. *Nat Genet* **34**, 421–428 (2003).

51. Ebermann, I. *et al.* A novel gene for Usher syndrome type 2: mutations in the long isoform of whirlin are associated with retinitis pigmentosa and sensorineural hearing loss. *Hum Genet* **121**, 203–211 (2007).

52. Delhommel, F. *et al.* Structural Characterization of Whirlin Reveals an Unexpected and Dynamic Supramodule Conformation of Its PDZ Tandem. *Structure* **25**, 1645-1656.e5 (2017).

53. Delhommel, F., Wolff, N. & Cordier, F. 1H, 13C and 15N backbone resonance assignments and dynamic properties of the PDZ tandem of Whirlin. *Biomol NMR Assign* **10**, 361–365 (2016).

54. Park, T. Crk and CrkL as Therapeutic Targets for Cancer Treatment. *Cells* **10**, 739 (2021).

55. Nichols, G. L. *et al.* Identification of CRKL as the constitutively phosphorylated 39-kD tyrosine phosphoprotein in chronic myelogenous leukemia cells. *Blood* **84**, 2912–2918 (1994).

56. Songyang, Z. & Cantley, L. C. SH2 domain specificity determination using oriented phosphopeptide library. *Methods Enzymol* **254**, 523–535 (1995).

57. Marasco, M. & Carlomagno, T. Specificity and regulation of phosphotyrosine signaling through SH2 domains. *J Struct Biol X* **4**, 100026 (2020).

58. Kaneko, T. *et al.* Loops govern SH2 domain specificity by controlling access to binding pockets. *Sci Signal* **3**, ra34 (2010).

59. Ren, R., Mayer, B. J., Cicchetti, P. & Baltimore, D. Identification of a ten-amino acid proline-rich SH3 binding site. *Science* **259**, 1157–1161 (1993).

60. Yu, H. *et al.* Structural basis for the binding of proline-rich peptides to SH3 domains. *Cell* **76**, 933–945 (1994).

61. Lim, W. A., Richards, F. M. & Fox, R. O. Structural determinants of peptide-binding orientation and of sequence specificity in SH3 domains. *Nature* **372**, 375–379 (1994).

62. Birge, R. B., Kalodimos, C., Inagaki, F. & Tanaka, S. Crk and CrkL adaptor proteins: networks for physiological and pathological signaling. *Cell Commun Signal* **7**, 13 (2009).

63. Perutz, M. F. Hemoglobin structure and respiratory transport. *Sci Am* **239**, 92–125 (1978).

64. Malagrinò, F. *et al.* On the Effects of Disordered Tails, Supertertiary Structure and Quinary Interactions on the Folding and Function of Protein Domains. *Biomolecules* **12**, 209 (2022).

65. Laursen, L., Kliche, J., Gianni, S. & Jemth, P. Supertertiary protein structure affects an allosteric network. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 24294–24304 (2020).

66. Buljan, M. & Bateman, A. The evolution of protein domain families. *Biochem Soc Trans* **37**, 751–755 (2009).

67. Benz, C. *et al.* Proteome-scale mapping of binding sites in the unstructured regions of the human proteome. *Mol Syst Biol* **18**, e10584 (2022).

68. Kruse, T. *et al.* Large scale discovery of coronavirus-host factor protein interaction motifs reveals SARS-CoV-2 specific mechanisms and vulnerabilities. *Nat Commun* **12**, 6761 (2021).

69. Fersht, A. R. & Sato, S. Phi-value analysis and the nature of protein-folding transition states. *Proc Natl Acad Sci U S A* **101**, 7976–7981 (2004).

70. Gianni, S. & Jemth, P. Conserved nucleation sites reinforce the significance of Phi value analysis in protein-folding studies. *IUBMB Life* **66**, 449–452 (2014).

71. Malagrinò, F. *et al.* Understanding the Binding Induced Folding of Intrinsically Disordered Proteins by Protein Engineering: Caveats and Pitfalls. *Int J Mol Sci* **21**, E3484 (2020).

72. Ivarsson, Y. *et al.* An on-pathway intermediate in the folding of a PDZ domain. *J Biol Chem* **282**, 8568–8572 (2007).

73. Travaglini-Allocatelli, C. *et al.* Exploring the cytochrome c folding mechanism: cytochrome c552 from thermus thermophilus folds through an on-pathway intermediate. *J Biol Chem* **278**, 41136–41140 (2003).

74. Jenik, M. *et al.* Protein frustratometer: a tool to localize energetic frustration in protein molecules. *Nucleic Acids Res* **40**, W348-351 (2012).

75. Jaber Chehayeb, R. & Boggon, T. J. SH2 Domain Binding: Diverse FLVRs of Partnership. *Front Endocrinol (Lausanne)* **11**, 575220 (2020).

76. Rickles, R. J. *et al.* Identification of Src, Fyn, Lyn, PI3K and Abl SH3 domain ligands using phage display libraries. *EMBO J* **13**, 5598–5604 (1994).

77. Kärkkäinen, S. *et al.* Identification of preferred protein interactions by phage-display of the human Src homology-3 proteome. *EMBO Rep* **7**, 186–191 (2006).

78. Cheadle, C. *et al.* Identification of a Src SH3 domain binding motif by screening a random phage display library. *J Biol Chem* **269**, 24034–24039 (1994).

79. Pagano, L., Malagrinò, F., Nardella, C., Gianni, S. & Toto, A. Experimental Characterization of the Interaction between the N-Terminal SH3 Domain of Crkl and C3G. *IJMS* **22**, 13174 (2021).

80. Posern, G. *et al.* Development of highly selective SH3 binding peptides for Crk and CRKL which disrupt Crk-complexes with DOCK180, SoS and C3G. *Oncogene* **16**, 1903–1912 (1998).

81. Oughtred, R. *et al.* The BioGRID interaction database: 2019 update. *Nucleic Acids Res* **47**, D529–D541 (2019).

82. Xie, M. *et al.* FAM110B Inhibits Non-Small Cell Lung Cancer Cell Proliferation and Invasion Through Inactivating Wnt/β-Catenin Signaling. *OTT* **Volume 13**, 4373–4384 (2020).

83. Tang, Z.-Z., Wang, H.-B., Zeng, M., Liu, C. & Li, D.-H. [DNA Repair Function and Mutation of an H2B Monoubiquitination Factor WDR70 in Ovarian Cancer]. *Sichuan Da Xue Xue Bao Yi Xue Ban* **48**, 693–698 (2017).

84. Dubrow, A., Lin, S., Savage, N., Shen, Q. & Cho, J.-H. Molecular Basis of the Ternary Interaction between NS1 of the 1918 Influenza A Virus, PI3K, and CRK. *Viruses* **12**, E338 (2020).

85. Jackson, S. E. How do small single-domain proteins fold? *Folding and Design* **3**, R81–R91 (1998).

86. Davey, N. E., Simonetti, L. & Ivarsson, Y. ProP-PD for proteome-wide motif-mediated interaction discovery. *Trends Biochem Sci* **47**, 547–548 (2022).

6. Scientific communications

1) **Pagano, L.**, Gkartziou, F., Aiello, S., Simonis, B., Ceccacci, F., Sennato, S., Ciogli, A., Mourtas, S.,Spiliopoulou, I., Antimisiaris, S.G., Bombelli, C., Mancini, G. Resveratrol loaded in cationic glucosylated liposomes to treat Staphylococcus epidermidis infections (2022) Chemistry and Physics of Lipids, 243, art. no. 105174.

2) Aiello, S., **Pagano, L.,** Ceccacci, F., Simonis, B., Sennato, S., Bugli, F., Martini, C., Torelli, R.,Sanguinetti, M., Ciogli, A., Bombelli, C., Mancini, G. Mannosyl, glucosyl or galactosyl liposomes to improve resveratrol efficacy against MethicillinResistant Staphylococcus aureus biofilm (2021) Colloids and Surfaces A: Physicochemical and Engineering Aspects, 617, art. no. 126321.

3) Nora, G.-I., Venkatasubramanian, R., Strindberg, S., Siqueira-Jørgensen, S.D., **Pagano, L.**, Romanski, F.S., Swarnakar, N.K., Rades, T., Müllertz, A. Combining lipid based drug delivery and amorphous solid dispersions for improved oral drugabsorption of a poorly water-soluble drug (2022) Journal of Controlled Release, 349, pp. 206-212.

4) **Pagano, L.,** Toto, A., Malagrinò, F., Visconti, L., Jemth, P., Gianni, S. Double mutant cycles as a tool to address folding, binding, and allostery (2021) International Journal of Molecular Sciences, 22 (2), art. no. 828, pp. 1-10.

5) **Pagano, L.**, Malagrinò, F., Visconti, L., Troilo, F., Pennacchietti, V., Nardella, C., Toto, A., Gianni, S. Probing the Effects of Local Frustration in the Folding of a Multidomain Protein (2021) Journal of Molecular Biology, 433 (15), art. no. 167087.

6) **Pagano, L.**, Malagrinò, F., Nardella, C., Gianni, S., Toto, A. Experimental characterization of the interaction between the n-terminal sh3 domain of crkl and c3g (2021) International Journal of Molecular Sciences, 22 (24), art. no. 13174.

7) Malagrinò, F., Fusco, G., Pennacchietti, V., Toto, A., Nardella, C., **Pagano, L.**, de Simone, A., Gianni,S. Cryptic binding properties of a transient folding intermediate in a PDZ tandem repeat (2022) Protein Science, 31 (9), art. no. 4396.

8) Nardella, C., Toto, A., Santorelli, D., **Pagano, L.**, Diop, A., Pennacchietti, V., Pietrangeli, P., Marcocci,L., Malagrinò, F., Gianni, S. Folding and Binding Mechanisms of the SH2 Domain from Crkl (2022) Biomolecules, 12 (8), art. no. 1014.

9) Toto, A., Malagrinò, F., Nardella, C., Pennacchietti, V., **Pagano, L.,** Santorelli, D., Diop, A., Gianni, S. Characterization of early and late transition states of the folding pathway of a SH2 domain (2022) Protein Science, 31 (6), art. no. 4332.

10) Malagrinò, F., Pennacchietti, V., Santorelli, D., **Pagano, L**., Nardella, C., Diop, A., Toto, A., Gianni, S. On the Effects of Disordered Tails, Supertertiary Structure and Quinary Interactions on the Foldingand Function of Protein Domains (2022) Biomolecules, 12 (2), art. no. 209.

11) Malagrinò, F., Diop, A., **Pagano, L**., Nardella, C., Toto, A., Gianni, S. Unveiling induced folding of intrinsically disordered proteins – Protein engineering, frustration and emerging themes (2022) Current Opinion in Structural Biology, 72, pp. 153-160.

12) Nardella, C., Visconti, L., Malagrinò, F., **Pagano, L**., Bufano, M., Nalli, M., Coluccia, A., La Regina, G.,Silvestri, R., Gianni, S., Toto, A. Targeting PDZ domains as potential treatment for viral infections, neurodegeneration and cancer (2021) Biology Direct, 16 (1), art. no. 15.

13) Nardella, C., Malagrinò, F., **Pagano, L**., Rinaldo, S., Gianni, S., Toto, A. Determining folding and binding properties of the C-terminal SH2 domain of SHP2 (2021) Protein Science, 30 (12), pp. 2385-2395.

14) Visconti, L., Malagrinò, F., Troilo, F., **Pagano, L**., Toto, A., Gianni, S. Folding and Misfolding of a PDZ Tandem Repeat: Folding of a PDZ tandem (2021) Journal of Molecular Biology, 433 (7), art. no. 166862.

15) Toto, A., Ma, S., Malagrinò, F., Visconti, L., **Pagano, L.**, Stromgaard, K., Gianni, S. Comparing the binding properties of peptides mimicking the Envelope protein of SARS-CoV andSARS-CoV-2 to the PDZ domain of the tight junction-associated PALS1 protein (2020) Protein Science, 29 (10), pp. 2038-2042.

16) Visconti, L., Malagrinò, F., **Pagano, L.**, Toto, A. Understanding the mechanism of recognition of gab2 by the N-SH2 domain of SHP2 (2020) Life, 10 (6), art. no. 85, pp. 1-9.

17) Toto, A., Malagrinò, F., Visconti, L., Troilo, F., **Pagano, L.**, Brunori, M., Jemth, P., Gianni, S. Templated folding of intrinsically disordered proteins (2020) Journal of Biological Chemistry, 295 (19), pp. 6586-6593.

18) Malagrinò, F., Visconti, L., **Pagano, L.,** Toto, A., Troilo, F., Gianni, S. Understanding the binding induced folding of intrinsically disordered proteins by protein engineering:Caveats and pitfalls (2020) International Journal of Molecular Sciences, 21 (10), art. no. 3484.

7. Attachments

Paper 1: Probing the effect of local frustration in the folding of a multidomain protein

Research Article

Probing the Effects of Local Frustration in the Folding of a Multidomain Protein

Livia Pagano, Francesca Malagrinò, Lorenzo Visconti, Francesca Troilo, Valeria Pennacchietti, Caterina Nardella, Angelo Toto * and Stefano Gianni *

Istituto Pasteur – Fondazione Cenci Bolognetti, *Dipartimento di Scienze Biochimiche "A. Rossi Fanelli" and Istituto di Biologia e Patologia Molecolari del CNR, Sapienza Università di Roma, 00185 Rome, Italy*

Correspondence to Angelo Toto and Stefano Gianni: *angelo.toto@uniroma.it (A. Toto), Stefano.gianni@uniroma1.it (S. Gianni)*
https://doi.org/10.1016/j.jmb.2021.167087
Edited by Daniel Otzen

Abstract

Our current knowledge of protein folding is primarily based on experimental data obtained from isolated domains. In fact, because of their complexity, multidomain proteins have been elusive to the experimental characterization. Thus, the folding of a domain in isolation is generally assumed to resemble what should be observed for more complex structural architectures. Here we compared the folding mechanism of a protein domain in isolation and in the context of its supramodular multidomain structure. By carrying out an extensive mutational analysis we illustrate that while the early events of folding are malleable and influenced by the absence/presence of the neighboring structures, the late events appear to be more robust. These effects may be explained by analyzing the local frustration patterns of the domain, providing critical support for the funneled energy landscape theory of protein folding, and highlighting the role of protein frustration in sculpting the early events of the reaction.

Introduction

The study of the mechanisms of folding of proteins has played a pivotal role in Molecular Biology over the past decades. By considering that correct folding is vital for essentially all cellular functions, it is of critical importance to analyze the mechanisms and forces that balance this process. In this context, a particularly elusive aspect has been represented by the inter-domain communication within multi-domain systems. In fact, despite two-thirds of eukaryotic proteins are formed by more than one domain, because of their complexity, only a few folding studies have been carried out on multi-domain proteins; our current knowledge being primarily based on works on single-domain systems.

A protein domain is generally defined as an independent sub-structure with specific equilibrium and kinetic properties.[1] Consequently, it is implicitly assumed that protein domains may fold independently and their folding in isolation is expected to resemble closely what should be observed in the context of more complex structural architectures. Nevertheless, the experimental works addressing these issues are relatively scarce and multi-domain folding has been addressed only in relatively few studies.[2–9] Importantly, while in some cases it has been concluded that adjacent domains do not affect their respective folding pathways,[10] in other cases the folding of multi-domain proteins has been found to compete with the accumulation of transient kinetic traps.[4,5,8,11–18] Haran and co-owrkers recently successfully manipulated the

folding landscape of a multidomain protein,[19] demonstrating a remarkable plasticity of domain folding, which may be modulated by rational site-directed mutagenesis. These findings highlight that the general grounds of multi-domain folding has yet to be elucidated and stress the importance of additional research.

Whirlin is a scaffolding protein expressed in both hair cells and retinal photoreceptor cells.[20–22] Among its complex structural architecture, it displays an N-terminal tandem comprising two PDZ domains (P1-P2) that are critical for the recognition and binding of other proteins of the hearing apparatus. From a structural perspective, NMR characterization of the tandem suggested the two PDZ populate transient supramodular architectures,[23–24] which improve the binding capacity of the first domain. We recently characterized the folding kinetics of P1-P2 and demonstrated the presence of a misfolded intermediate that competes with productive folding,[11,12] a finding that parallels earlier observations on engineered WW domain tandems.[25] Moreover, the fortuitous differences in stability between PDZ1 and PDZ2 allowed us to show that only the concurrent denaturation of both domains leads to the accumulation of the misfolded kinetic trap that competes with native folding and substantially slows the productive pathway. Whereas the misfolded trap cannot populate when PDZ2 is held in its native conformation.

Whilst it is clear that the presence of denatured PDZ2 has a pronounced effect on the folding of PDZ1, is there any effect when the former is held native? How robust is the folding of PDZ1? In the specific effort to address these questions, we report here a complete mutational analysis of the folding pathway of P1-P2, in comparison to that of PDZ1 expressed in isolation. Thanks to the analysis of the folding and unfolding kinetics of 52 site directed mutants, we successfully provide a glimpse of the robustness of the folding of this protein domain in the context of a multi-domain construct. In particular, we show that whilst the late events of folding appear insensitive to the presence of a contiguous folded domain, the early events are more malleable and display some detectable deviations. Strikingly, these deviations may be well explained by analyzing the local frustration patterns that may be calculated for PDZ1, reinforcing the importance of the funneled energy landscape theory in protein folding.[26]

Results

Comparing the folding kinetics of PDZ1 and P1-P2

The main purpose of this study lies in comparing the folding of a protein domain, PDZ1 from Whirlin, in isolation and in the context of its multi-domain supramodular organization, P1-P2. To achieve this aim, we resorted to perform a complete kinetic analysis of the folding and unfolding properties of these two constructs. In analogy to our previous works,[11,12] in all the proteins described in this study, we inserted a mutation, Y168W, whose fluorescence reports efficiently the denaturation of PDZ1, whereas no changes in fluorescence can be observed in the case of PDZ2.

The folding and unfolding kinetics of PDZ1 was investigated by stopped-flow experiments. In all cases, folding and unfolding time courses were fitted satisfactorily to a single exponential decay at

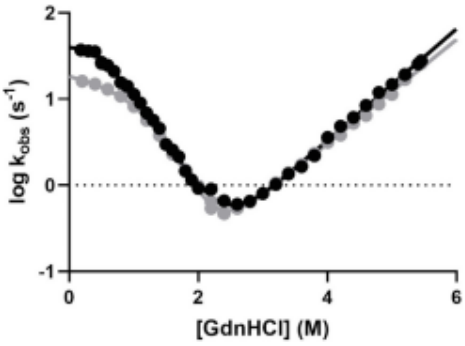


Figure 1. Semi-logarithmic plot of the observed folding/unfolding rate constant versus denaturant concentration. Chevron plot of PDZ1 (black circles) and P1P2 (gray circles) both fitted with a three-state equation, which indicates the presence of a folding intermediate. As detailed in the text, the P1P2 refolding branch was obtained with the protein diluted in mild denaturant concentrations (i.e., 2.2 M GdnHCl) to ensure PDZ2 domain of the tandem to preserve a folded conformation.

Table 1 Kinetic and thermodynamic (un)folding parameters of PDZ1 and P1P2 site-directed variants

	k_{N} (s^{-1})	k_{NI} (s^{-1} 10^{-3})	K_{DI} (10^{-6})	$\Delta\Delta G_{DI}$ (kcal mol^{-1})	$\Delta\Delta G_{TS2-N}$ (kcal mol^{-1})	$\Delta\Delta G_{D-N}$ (kcal mol^{-1})	Φ_{I}	Φ_{TS2}
PDZ1 WT	43 ± 2	15 ± 1	45 ± 10					
PDZ1 L141A	57 ± 5	120 ± 10	400 ± 80	1.30 ± 0.10	1.23 ± 0.03	2.40 ± 0.20	0.55 ± 0.07	0.48 ± 0.04
PDZ1 A147G	56 ± 3	32 ± 1	100 ± 10	0.50 ± 0.10	0.44 ± 0.03	0.80 ± 0.10	0.63 ± 0.16	0.44 ± 0.09
PDZ1 L153A	31 ± 3	47 ± 2	370 ± 70	1.20 ± 0.10	0.66 ± 0.04	2.10 ± 0.20	0.59 ± 0.08	0.69 ± 0.03
PDZ1 I157V	48 ± 3	44 ± 2	80 ± 10	0.30 ± 0.10	0.63 ± 0.04	0.90 ± 0.10	0.36 ± 0.15	0.30 ± 0.12
PDZ1V165A	43 ± 2	20 ± 1	60 ± 10	0.20 ± 0.10	0.15 ± 0.03	0.30 ± 0.10	/	/
PDZ1 I167V	32 ± 1	30 ± 1	80 ± 10	0.30 ± 0.10	0.39 ± 0.03	0.90 ± 0.10	0.35 ± 0.12	0.56 ± 0.07
PDZ1 V169A	5 ± 1	57 ± 4	270 ± 130	1.00 ± 0.30	0.78 ± 0.05	3.10 ± 0.30	0.34 ± 0.10	0.75 ± 0.03
PDZ1 L171A	97 ± 5	13 ± 1	27 ± 3	−0.30 ± 0.10	−0.11 ± 0.04	−0.90 ± 0.10	0.34 ± 0.13	0.87 ± 0.04
PDZ1 L177A	38 ± 2	20 ± 1	75 ± 10	0.30 ± 0.10	0.16 ± 0.03	0.50 ± 0.10	0.56 ± 0.23	0.70 ± 0.09
PDZ1 A178G	15 ± 1	90 ± 2	290 ± 30	1.10 ± 0.10	1.04 ± 0.03	2.80 ± 0.10	0.40 ± 0.04	0.62 ± 0.02
PDZ1 V185A	34 ± 2	38 ± 2	100 ± 20	0.50 ± 0.10	0.54 ± 0.04	1.20 ± 0.10	0.42 ± 0.12	0.54 ± 0.06
PDZ1 I189V	28 ± 1	50 ± 1	100 ± 10	0.50 ± 0.10	0.70 ± 0.03	1.40 ± 0.10	0.33 ± 0.08	0.51 ± 0.04
PDZ1 L190A	33 ± 2	74 ± 2	230 ± 40	1.00 ± 0.10	0.93 ± 0.03	2.10 ± 0.10	0.47 ± 0.07	0.55 ± 0.03
PDZ1 V192A	29 ± 1	34 ± 1	200 ± 20	0.90 ± 0.10	0.47 ± 0.03	1.60 ± 0.10	0.55 ± 0.08	0.70 ± 0.03
PDZ1 L197A	34 ± 2	290 ± 10	270 ± 60	1.10 ± 0.10	1.74 ± 0.03	2.90 ± 0.20	0.36 ± 0.06	0.41 ± 0.03
PDZ1 A198G	47 ± 2	43 ± 1	80 ± 10	0.30 ± 0.10	0.61 ± 0.03	0.90 ± 0.10	0.37 ± 0.13	0.32 ± 0.09
PDZ1 A203G	38 ± 3	36 ± 2	120 ± 20	0.60 ± 0.10	0.51 ± 0.04	1.20 ± 0.20	0.50 ± 0.13	0.57 ± 0.06
PDZ1 A205G	26 ± 5	140 ± 10	400 ± 210	1.30 ± 0.30	1.31 ± 0.05	2.90 ± 0.40	0.44 ± 0.12	0.55 ± 0.06
PDZ1 V206A	39 ± 5	42 ± 3	260 ± 70	1.00 ± 0.20	0.59 ± 0.04	1.70 ± 0.20	0.61 ± 0.13	0.65 ± 0.05
PDZ1 A208G	28 ± 7	280 ± 20	1100 ± 600	1.90 ± 0.30	1.71 ± 0.05	3.80 ± 0.40	0.49 ± 0.10	0.56 ± 0.05
PDZ1 L215A	26 ± 4	68 ± 5	260 ± 100	1.00 ± 0.20	0.88 ± 0.05	2.20 ± 0.30	0.47 ± 0.12	0.60 ± 0.05
PDZ1 V216A	28 ± 3	57 ± 3	200 ± 40	0.90 ± 0.10	0.78 ± 0.04	1.80 ± 0.20	0.51 ± 0.09	0.56 ± 0.05
P1P2 WT	23 ± 1	13 ± 1	16 ± 1					
P1P2 L141A	23 ± 2	63 ± 4	20 ± 10	0.20 ± 0.30	0.92 ± 0.04	1.20 ± 0.30	0.20 ± 0.22	0.21 ± 0.18
P1P2 A147G	16 ± 1	18 ± 1	10 ± 2	−0.30 ± 0.10	0.19 ± 0.03	0.10 ± 0.10	/	/
P1P2 L153A	9 ± 1	31 ± 2	20 ± 10	0.10 ± 0.30	0.50 ± 0.04	1.10 ± 0.30	/	/
P1P2 I157V	11 ± 1	23 ± 2	10 ± 4	−0.30 ± 0.30	0.32 ± 0.05	0.50 ± 0.30	−0.68 ± 0.76	0.29 ± 0.47
P1P1 V165A	17 ± 1	16 ± 1	10 ± 2	−0.20 ± 0.10	0.13 ± 0.03	0.10 ± 0.10	/	/
P1P2 I167V	16 ± 1	21 ± 1	10 ± 2	−0.20 ± 0.10	0.29 ± 0.03	0.30 ± 0.20	/	/
P1P2 V169A	2 ± 1	59 ± 3	130 ± 70	1.30 ± 0.30	0.89 ± 0.03	3.50 ± 0.30	0.35 ± 0.09	0.75 ± 0.02
P1P2 L171A	58 ± 5	10 ± 1	10 ± 3	−0.30 ± 0.20	−0.17 ± 0.05	−1.00 ± 0.20	0.28 ± 0.20	0.83 ± 0.06
P1P2 L177A	15 ± 1	12 ± 1	40 ± 10	0.60 ± 0.10	−0.03 ± 0.03	0.80 ± 0.20	0.74 ± 0.22	1.04 ± 0.04
P1P2 A178G	7 ± 1	100 ± 5	400 ± 160	1.90 ± 0.30	1.21 ± 0.03	3.80 ± 0.30	0.50 ± 0.07	0.68 ± 0.02
P1P2 V185A	13 ± 1	30 ± 1	50 ± 10	0.70 ± 0.20	0.48 ± 0.03	1.50 ± 0.20	0.47 ± 0.11	0.69 ± 0.04
P1P2 I189V	11 ± 1	35 ± 3	10 ± 10	−0.20 ± 0.30	0.58 ± 0.04	0.90 ± 0.30	−0.20 ± 0.40	0.30 ± 0.30
P1P2 L190A	15 ± 1	58 ± 2	80 ± 10	1.00 ± 0.10	0.88 ± 0.02	2.10 ± 0.10	0.47 ± 0.06	0.58 ± 0.03

any final denaturant concentration. A semi-logarithmic plot of the observed folding/unfolding rate constant versus denaturant concentration (chevron plot) is reported in Figure 1. Whereas the logarithm of the unfolding rate constant appears to vary linearly with denaturant concentration, the refolding rate constant displays a deviation from linearity (roll-over effect), which indicates the presence of a folding intermediate. Accordingly, the dependence of the observed rate constant on denaturant concentration was fitted to a three-state equation as formalized in the Experimental Section. The calculated folding and unfolding parameters are reported in Table 1.

We previously showed that the equilibrium unfolding of P1-P2 proceeds in a stepwise manner and the first PDZ domain may be denatured at 'mild denaturant concentration', i.e. 2.2 M of GdnHCl, whereas the second PDZ domain retains its native conformation.[11,12] Thus, in an effort to compare the folding of PDZ1, in isolation and in the context of its multi-domain supramodular organization, we measured the chevron plot of P1-P2 by first denaturing the protein in 2.2 M GdnHCl and then triggering refolding by rapidly mixing with buffer in the presence of increasing concentrations of GdnHCl. Conversely, unfolding was measured by challenging native P1-P2 with GdnHCl at different concentrations. Also in this case, both folding and unfolding time courses were satisfactorily fitted to a single exponential decay at any final denaturant concentration. A superposition between the chevron plots of PDZ1 in isolation and in the tandem is reported in Figure 1. Whilst the unfolding of the two proteins appears essentially identical, it might be noticed that inclusion of PDZ1 in a tandem repeat results in a minor but detectable stabilization of the folding intermediate, as probed by the slightly more pronounced roll-over effect in the refolding branch. Importantly, this intermediate is distinct from the major misfolding events previously reported[11,12] that may occur only when both PDZ domains are unfolded in P1-P2 and cannot be detected in PDZ1 in isolation. In analogy to what reported above for PDZ1, data were fitted to a three state folding equation and the resulting parameters are listed in Table 1.

The folding mechanism of PDZ1 – Φ value analysis

To address the details of the folding of PDZ1 we performed Φ value analysis. By following this methodology, residue-specific structural information of metastable state(s) along the reaction pathway is inferred by comparing the

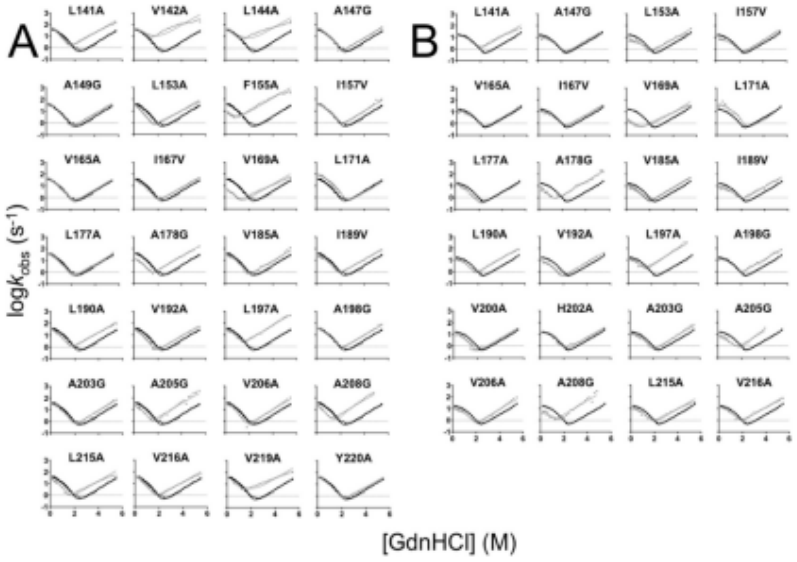


Figure 2. Chevron plots of the 52 site-directed mutants. The PDZ1 wt (Panel A) and the P1P2 wt (Panel B) are reported in black and their mutated variants in gray. Data were fitted with Eq. (1), as formalized in Material and Methods section. The buffer used in all the experiments was 50 mM TrisHCl pH 7.5 and 0.3 M Na_2SO_4 and the temperature was 25 °C.

kinetics of folding of the wild-type protein with those of a series of conservative single mutants.[27,28] Quantitatively, the strength of the contacts is measured by the Φ value, which normalizes the stability change of the metastable state upon mutation to that of the native state. A Φ value close to 1 is indicative of native-like structure in the metastable state of that specific residue, whereas a Φ value equal to 0 suggests that the mutated residue is as unstructured in the metastable state as it is in the denatured state.

Twenty-eight site-directed variants of PDZ1 were produced, expressed and subjected to folding and unfolding experiments. The mutants were designed according to standard rules of Φ value analysis, which were extensively discussed elsewhere.[27–30] In summary, a conservative deletion of hydrophobic side chains was designed, a type of mutation that represents the easiest to be interpreted. The chevron plot obtained for each variant is reported in Figure 2.

In analogy to previously works on multistate systems,[31–33] the obtained chevron plots were globally fitted to a three-state equation with shared m-values. The resulting folding parameters are listed in Table 1.

To provide structural information of the intermediate and transition states, the mutants were divided in three groups based on their measured Φ values: small (Φ < 0.3; red), intermediate (0.3 < Φ < 0.7; magenta), and large (Φ > 0.7; blue). The color-coded mutations were then mapped onto the structure of PDZ1 (Figure 3). The structural distribution of the measured values suggests the intermediate to be characterized by the embryonic formation of a weak nucleus, which appears to be diffused in the whole globule. Conversely, native structure in the transition state presents a more polarized distribution, encompassing mostly the residues found in the β3-β4 strands.

Testing the robustness of single domain folding – Φ value analysis of PDZ1 in the P1-P2 construct

To test the robustness of the folding of PDZ1 in the context of the mutidomain supramodular structure we resorted to perform a comparative Φ value analysis. Hence, in analogy to what described above, we expressed, purified and characterized twenty-four site directed mutants

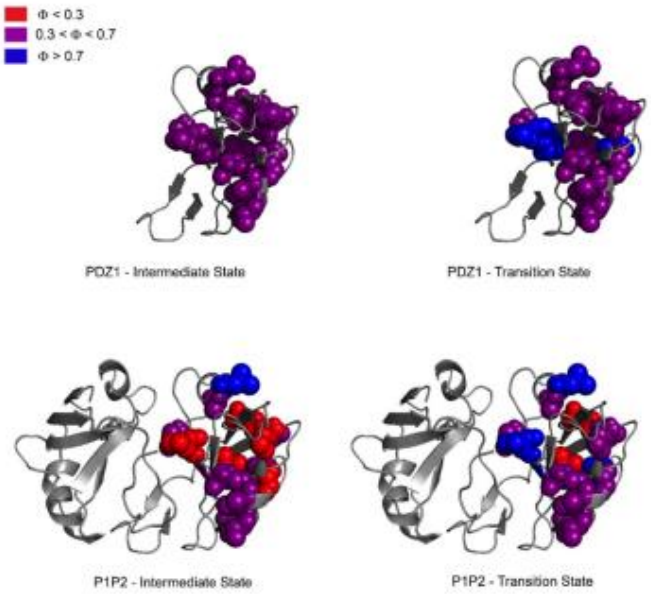


Figure 3. Φ values mapped on the PDZ1 and P1P2 3D structures. Color-coded distribution of Φ values is calculated for the intermediate state and transition state on the PDZ1 and P1P2 structures (Φ < 0.3; red) 0.3 < Φ < 0.7; magenta Φ > 0.7; blue). While the intermediate states appear to be more malleable and characterized by a lower degree of native-like structure in P1P2 compared to PDZ1, this difference is less pronounced in the comparison of transition states.

and subjected them to kinetic folding and unfolding experiments.

Figure 2 reports a mutant-by-mutant comparison of the chevron plots measured for each of the variant measured for PDZ1 in isolation and in the context of its supramodular architecture, with the calculated folding and unfolding parameters listed in Table 1. Overall, it might be observed that in essentially all cases, the chevron plots of all the variants appear nearly identical in both constructs, with the relevant exception of the refolding roll over that appears more pronounced in the case of the P1-P2 constructs rather than in the case of PDZ1 in isolation.

In the P1P2 construct, the architecture of the intermediate is less structured compared to the isolated domain, as shown by the presence of lower Φ -values. This finding contrasts what observed for the transition state of folding, which is rather robust and structurally similar to what observed in the case of PDZ1 in isolation. Thus, while in the late stages the two domains display the same robust pattern, in the early events the mutational analysis reveals significant differences between PDZ1 in isolation and in tandem.

A direct way to compare mutational data sets is to perform Φ-Φ plots of a relevant state, as well as in comparing the changes in free energies upon mutation.[29] Figure 4 depicts the Φ-Φ and ΔΔG plots for the intermediate and transitions states of PDZ1 in isolation and in the P1-P2 construct. It is evident that, whilst the data for the transition state are conserved in the two constructs, consistent with a linear correlation and a slope of 1, in the case of the intermediate there are clear differences between the two constructs. Thus, there is an intriguing picture emerging from the comparison between PDZ1 in isolation and in the context of its supramodular organization that suggests a clear robustness to characterize the late stages of folding, whereas a more malleable behavior may be detected in the early stages.

Discussion

The Levinthal paradox emphasizes how the mechanism of protein folding cannot take place via a stochastic search between all possible conformations.[34] One of the most elegant theories to disentangle the paradox is to postulate the free energy of proteins to be minimally frustrated.[26,35,36] By following this view, there is a strong energetic bias towards the native conformation and alternative structures display a very marginal stability, implying a funneled energy landscape. Nevertheless, it should be noted that proteins are not only optimized to fold but also to function. Thus, because the evolutionary demands for folding and function might contrast, it is observed that functional protein sites are often associated with patterns of local frustration of non-optimized sequences.[37–40] The effects of folding pathways of such frustration,

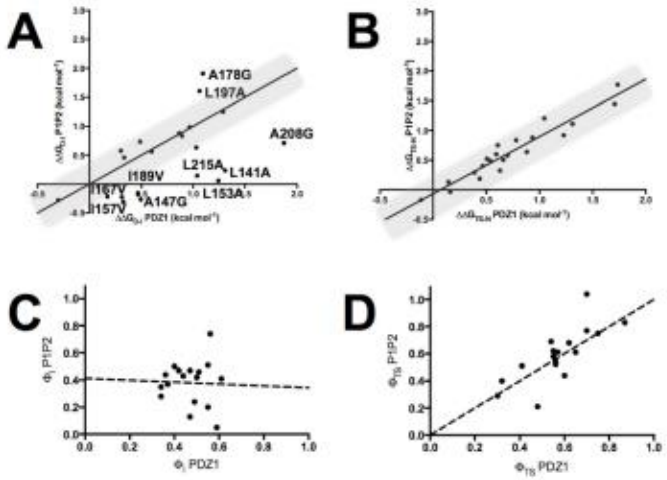


Figure 4. ΔΔG plots and Φ-Φ plots for PDZ1 and P1-P2 intermediate states (Panels A and C) and transitions states (Panels B and D). Each point in the graphs represents a single site-directed mutation occurring in both proteins. While for the intermediate states there is a pronounced scattered distribution, with several variants lying outside of the linear fit, a strong linear correlation is evident for the analysis of the transition states.

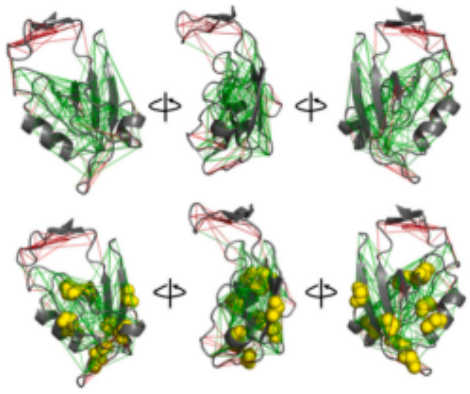


Figure 5. Frustration and folding of PDZ1. Frustration patterns were calculated from the algorithm kindly provided by Wolynes, Ferreiro and co-workers and available at the website http://frustratometer.qb.fcen.uba.ar .The red lines indicate local frustrated patterns within the structure. The yellow spheres are the residues highlighted in Figure 4(A) that show a detectable difference in $\Delta\Delta G_{D\text{-}I}$, when the folding of the domain is studied in isolation as compared to in the context of its supramodular structure.

which has been addressed only in a few cases in the case of single domain proteins,[41–44] are still poorly understood in more complex systems.

The experimental characterization of the folding pathway of PDZ1 from whirlin in isolation and in the context of its multidomain supramodular structure highlights that, whilst the late events appear strongly committed to the native topology, the early events are more malleable and prone to changes depending on the presence/absence of the adjacent domain. We note that such behavior parallels what expected from the funneled energy landscape theory that postulates the native bias to be weak at early stages of folding, allowing for alternative early folding events.

Ferreiro, Wolynes and co-workers provided a public algorithm that allows calculating the frustration patterns of proteins (available at http://frustratometer.qb.fcen.uba.ar).[45] On the light of what summarized above, to explain the observed differences in the folding intermediate of PDZ1 in isolation and in the context of its multidomain supramodular structure, we calculated the frustration pattern of PDZ1. Figure 5 depicts the structure of PDZ1 and highlights the frustration patterns within this protein domain along with the residues that are prone to structural changes in the folding intermediate. Strikingly, we found a remarkable superposition between residues prone to alternative folding pathways and the frustrated regions of PDZ1. In particular, out of 21 comparable mutants between the isolated domain and the tandem repeat, we found 8 predicted true positives, 2 false positives, 3 false negatives and 8 true negatives. This finding further confirms that frustration sculpts the early stages of folding, whereas it has little effects on the late stages of the reaction.

Conclusions

Despite decades of intense research, our understanding on the folding of multidomain proteins is still relatively limited. Moreover, it may be noted that general rules describing the folding of multidomain systems appear rather unlikely. In fact, whilst in some cases protein domains appear to behave as bead-on-a-string and each individual domain is capable to fold and function in the absence of the remainder of the protein, in other cases, the multidomain architecture is critical to guarantee the stability of each domain. By considering the mechanisms of folding, it has been shown that, in general, a complex architecture results in increasing the repertoire of available pathways, with multidomain proteins typically displaying parallel folding routes.[14–19] Notably, such complexity may results in the transient accumulation of misfolded conformations that compete with productive folding.[4,5,12,13]

By considering the above observations, the results presented in this work provide additional clues in understanding multi-domain folding. In fact, by providing a detailed comparison between the folding pathway of a protein domain in isolation and in the context of its supramodular multidomain structure, we show that whilst the late events of folding are robust, the early events are much more malleable and prone to structural changes. Remarkably, such changes parallel the

frustration patterns of the domain, indicating that, in multi-domain folding, alternative pathways, and possibly protein misfolding, may arise from these locally non-optimized regions. Whilst these observations appear in line with previous characterization of multidomain systems, we wish to notice that no previous study correlated transient misfolding with local frustration patterns, thus providing additional clues to the current understanding of multidomain systems. Furthermore, it is worth noticing that the results presented in this work represent an experimental validation of the funneled energy landscape in protein folding and open a new interesting view in the interpretation of the observed complexity of the folding of multidomain proteins. Future work on other systems will further confirm the generality of these observations.

Materials and Methods

Expression and purification of PDZ1 and P1P2 and site-directed variants

PDZ1 and tandem P1-P2 proteins were expressed and purified as previously described.[12] Site-directed variants were obtained using the Quik-Change Lightning Mutagenesis Kit (Agilent) following manufacturer instructions.

Kinetic folding experiments

Rapid-mixing kinetic folding and unfolding experiments were carried out with SX-18 and PiStar stopped-flow devices (Applied Photophysics). For all the experiments, the excitation wavelength was 280 nm and fluorescence emissions were measured with a 320-nm cut-off filter. Protein final protein concentration was 1.5 μM. The temperature was set at 25 °C and the buffer used was 50 mM TrisHCl pH 7.5 and 0.3 M Na_2SO_4. Refolding experiments of P1P2 were performed with the protein diluted in mild denaturant concentration (i.e. 2.2 M GdnHCl). For each experiment, an average calculated from at least 5 independent traces was satisfactorily fitted with a single exponential equation. Semilogarithmic plot of observed rate constants versus [GdnHCl] (i.e., chevron plot) was fitted using the following equation

$$k_{obs} = \frac{\left(k^0_{IN}\exp\left(\frac{-m_{IN}[\mathrm{GdnHCl}]}{RT}\right)\right)}{\left(1+K_{DI}\exp\left(\frac{m_{DI}[\mathrm{GdnHCl}]}{RT}\right)\right)} + k^0_{NI} \times \exp\left(\frac{m_{NI}[\mathrm{GdnHCl}]}{RT}\right) \quad (1)$$

Data obtained for PDZ1 variants and P1P2 variants were globally fitted with Prism software (GraphPad Software, Inc.) by sharing the m_{IN} m_{NI} and m_{DI} values for all data sets.

CRediT authorship contribution statement

Livia Pagano: Investigation, Formal analysis, Writing - review & editing. **Francesca Malagrinò:** Investigation, Formal analysis. **Lorenzo Visconti:** Investigation, Formal analysis. **Francesca Troilo:** Investigation, Formal analysis. **Valeria Pennacchietti:** Investigation, Formal analysis. **Caterina Nardella:** Investigation, Formal analysis. **Angelo Toto:** Investigation, Formal analysis, Writing - review & editing. **Stefano Gianni:** Conceptualization, Formal analysis, Supervision, Writing - original draft, Writing - review & editing.

Acknowledgments

Work partly supported by grants from the Italian Ministero dell'Istruzione dell'Universita' e della Ricerca (Progetto di Interesse 'Invecchiamento' to S.G.), Sapienza University of Rome (RP11715C34AEAC9B and RM1181641C2C24B9, RM11916B414C897E to S. G), the Associazione Italiana per la Ricerca sul Cancro (Individual Grant – IG 24551 to S.G.) the Istituto Pasteur Italia (Teresa Ariaudo Research Project 2018, to A.T.). F.M. was supported by a fellowship from the FIRC - Associazione Italiana per la Ricerca sul Cancro (Filomena Todini fellowship).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Received 10 April 2021;
Accepted 28 May 2021;
Available online 3 June 2021

Keywords:
PDZ;
ϕ-value analysis;
energy landscape;
tandem;
intermediate

References

1. Wetlaufer, D.B., (1973). Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci. USA*, **70**, 697–701.
2. Batey, S., Clarke, J., (2006). Apparent cooperativity in the folding of multidomain proteins depends on the relative rates of folding of the constituent domains. *Proc. Natl. Acad. Sci. USA*, **103**, 18113–18118.

3. Batey, S., Scott, K.A., Clarke, J., (2006). Complex folding kinetics of a multidomain protein. *Biophys. J.*, **90**, 2120–2130.
4. Borgia, A., Kemplen, K.R., Borgia, M.B., Soranno, A., Shammas, S., Wunderlich, B., Nettels, D., Best, R.B., et al., (2015). Transient misfolding dominates multidomain protein folding. *Nature Commun.*, **6**, 8861.
5. Borgia, M.B., Borgia, A., Best, R.B., Steward, A., Nettels, D., Wunderlich, B., Schuler, B., Clarke, J., (2011). Single-molecule fluorescence reveals sequence-specific misfolding in multidomain proteins. *Nature*, **474**, 662–665.
6. Han, J.H., Batey, S., Nickson, A.A., Teichmann, S.A., Clarke, J., (2007). The folding and evolution of multidomain proteins. *Nature Rev. Mol. Cell. Biol.*, **8**, 319–330.
7. Kumar, V., Chaudhuri, T.K., (2018). Spontaneous refolding of the large multidomain protein malate synthase G proceeds through misfolding traps. *J. Biol. Chem.*, **293**, 13270–13283.
8. Tian, P., Best, R.B., (2016). Structural determinants of misfolding in multidomain proteins. *PLoS Comput. Biol.*, **12**, e1004933
9. Arora, P., Hammes, G.G., Oas, T.G., (2006). Folding mechanism of a multiple independently-folding domain protein: double B domain of protein A. *Biochemistry*, **45**, 12312–12324.
10. Batey, S., Clarke, J., (2008). The folding pathway of a single domain in a multidomain protein is not affected by its neighbouring domain. *J. Mol. Biol.*, **378**, 297–301.
11. Visconti, L., Malagrinò, F., Troilo, F., Pagano, L., Toto, A., Gianni, S., (2021). Folding and misfolding of a PDZ tandem repeat. *J. Mol. Biol.*, **433**, 166862
12. Gautier, C., Troilo, F., Cordier, F., Malagrinò, F., Toto, A., Visconti, L., Zhu, Y., Brunori, M., et al., (2020). Hidden kinetic traps in multidomain folding highlight the presence of a misfolded but functionally competent intermediate. *Proc. Natl. Acad. Sci. USA*, **117**, 19963–19969.
13. Lafita, A., Tian, P., Best, R.B., Bateman, A., (2019). Tandem domain swapping: determinants of multidomain protein misfolding. *Curr. Opin. Struct. Biol.*, **58**, 97–104.
14. Beechem, J.M., Sherman, M.A., Mas, M.T., (1995). Sequential domain unfolding in phosphoglycerate kinase: fluorescence intensity and anisotropy stopped-flow kinetics of several tryptophan mutants. *Biochemistry*, **34**, 13943–13948.
15. Osváth, S., Köhler, G., Závodszky, P., Fidy, J., (1995). Asymmetric effect of domain interactions on the kinetics of folding in yeast phosphoglycerate kinase. *Protein Sci.*, **14**, 1609–1616.
16. Petersen, M., Barrick, D., (2021). Analysis of tandem repeat protein folding using nearest-neighbor models. *Annu. Rev. Biophys.*, **50**, 245–265.
17. Kloss, E., Barrick, D., (2008). Thermodynamics, kinetics, and salt dependence of folding of YopM, a large leucine-rich repeat protein. *J. Mol. Biol.*, **383**, 1195–1209.
18. Kloss, E., Courtemanche, N., Barrick, D., (2008). Repeat-protein folding: new insights into origins of cooperativity, stability, and topology. *Arch. Biochem. Biophys.*, **469**, 83–99.
19. Kantaev, R., Riven, I., Goldenzweig, A., Barak, Y., Dym, O., Peleg, Y., Albeck, S., Fleishman, S.J., et al., (2018). Manipulating the folding landscape of a multidomain protein. *J. Phys. Chem. B*, **122**, 11030–11038.
20. Ebermann, I., Scholl, H.P.N., Charbel Issa, P., Becirovic, E., Lamprecht, J., Jurklies, B., Millan, J.M., Aller, E., et al., (2007). A novel gene for Usher syndrome type 2: mutations in the long isoform of whirlin are associated with retinitis pigmentosa and sensorineural hearing loss. *Hum. Genet.*, **121**, 203–211.
21. Mathur, P.D., Yang, J., (2019). Usher syndrome and non-syndromic deafness: functions of different whirlin isoforms in the cochlea, vestibular organs, and retina. *Hear Res.*, **375**, 14–24.
22. Sorusch, N., Bauß, K., Plutniok, J., Samanta, A., Knapp, B., Nagel-Wolfrum, K., Wolfrum, U., (2017). Characterization of the ternary Usher syndrome SANS/ush2a/whirlin protein complex. *Hum. Mol. Genet.*, **26**, 1157–1172.
23. Delhommel, F., Cordier, F., Bardiaux, B., Bouvier, G., Colcombet-Cazenave, B., Brier, S., Raynal, B., Nouaille, S., et al., (2017). Structural characterization of Whirlin reveals an unexpected and dynamic supramodule conformation of its PDZ tandem. *Structure*, **25**, 1645–1656.
24. Delhommel, F., Wolff, N., Cordier, F., (2016). (1)H, (13)C and (15)N backbone resonance assignments and dynamic properties of the PDZ tandem of Whirlin. *Biomol. NMR Assign.*, **10**, 361–365.
25. Dave, K., Gasic, A.G., Cheung, M.S., Gruebele, M., (2019). Competition of individual domain folding with inter-domain interaction in WW domain engineered repeat proteins. *Phys. Chem. Chem. Phys.*, **21**, 24393–24405.
26. Bryngelson, J.D., Onuchic, J.N., Socci, N.D., Wolynes, P. G., (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins*, **21**, 167–195.
27. Fersht, A.R., Matouschek, A., Serrano, L., (1992). The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.*, **224**, 771–782.
28. Fersht, A.R., Sato, S., (2004). Phi-value analysis and the nature of protein-folding transition states. *Proc. Natl. Acad. Sci. USA*, **101**, 7976–7981.
29. Gianni, S., Jemth, P., (2014). Conserved nucleation sites reinforce the significance of Phi value analysis in protein-folding studies. *IUBMB Life*, **66**, 449–452.
30. Malagrinò, F., Visconti, L., Pagano, L., Toto, A., Troilo, F., Gianni, S., (2020). Understanding the binding induced folding of intrinsically disordered proteins by protein engineering: caveats and pitfalls. *Int. J. Mol. Sci.*, **21**, 3484.
31. Parker, M.J., Spencer, J., Clarke, A.R., (1995). An integrated kinetic analysis of intermediates and transition states in protein folding reactions. *J. Mol. Biol.*, **253**, 771–786.
32. Ivarsson, Y., Travaglini-Allocatelli, C., Jemth, P., Malatesta, F., Brunori, M., Gianni, S., (2007). An on-pathway intermediate in the folding of a PDZ domain. *J. Biol. Chem.*, **282**, 8568–8572.
33. Travaglini-Allocatelli, C., Gianni, S., Morea, V., Tramontano, A., Soulimane, T., Brunori, M., (2003). Exploring the cytochrome c folding mechanism: cytochrome c552 from thermus thermophilus folds through an on-pathway intermediate. *J. Biol. Chem.*, **278**, 41136–41140.
34. Levinthal, C., (1968). Are there pathways for protein folding?. *J. Chem. Phys.*, **65**, 44–45.
35. Onuchic, J.N., Socci, N.D., Luthey-Schulten, Z., Wolynes, P.G., (1996). Protein folding funnels: the nature of the transition state ensemble. *Fold. Des.*, **1**, 441–450.
36. Wolynes, P.G., (2005). Energy landscapes and solved protein-folding problems. *Philos. Transact. Roy. Soc. A Math. Phys. Eng. Sci.*, **363**, 453–464.

37. Ferreiro, D.U., Hegler, J.A., Komives, E.A., Wolynes, P.G., (2007). Localizing frustration in native proteins and protein assemblies. *Proc. Natl. Acad. Sci. USA*, **104**, 19819–19824.
38. Ferreiro, D.U., Hegler, J.A., Komives, E.A., Wolynes, P.G., (2011). On the role of frustration in the energy landscapes of allosteric proteins. *Proc. Natl. Acad. Sci. USA*, **108**, 3499–3503.
39. Sutto, L., Lätzer, J., Hegler, J.A., Ferreiro, D.U., Wolynes, P.G., (2007). Consequences of localized frustration for the folding mechanism of the IM7 protein. *Proc. Natl. Acad. Sci.*, **104**, 19825–19830.
40. Gianni, S., Freiberger, M.I., Jemth, P., Ferreiro, D.U., Wolynes, P.G., Fuxreiter, M., (2021). Fuzziness and frustration in the energy landscape of protein folding, function, and assembly. *Acc. Chem. Res.*, **54**, 1251–1259.
41. Di Silvio, E., Brunori, M., Gianni, S., (2015). Frustration sculpts the early stages of protein folding. *Angew. Chem. Int. Ed.*, **54**, 10867–10869.
42. Gianni, S., Camilloni, C., Giri, R., Toto, A., Bonetti, D., Morrone, A., Sormanni, P., Brunori, M., Vendruscolo, M., (2014). Understanding the frustration arising from the competition between function, misfolding, and aggregation in a globular protein. *Proc. Natl. Acad. Sci. USA*, **111**, 14141–14146.
43. Narayan, A., Gopi, S., Lukose, B., Naganathan, A.N., (2020). Electrostatic frustration shapes folding mechanistic differences in paralogous bacterial stress response proteins. *J. Mol. Biol.*, **432**, 4830–4839.
44. Halloran, K.T., Wang, Y., Arora, K., Chakravarthy, S., Irving, T.C., Bilsel, O., Brooks, C.L.R., Matthews, C.R., (2019). Frustration and folding of a TIM barrel protein. *Proc. Natl. Acad. Sci. USA*, **116**, 16378–16383.
45. Jenik, M., Parra, R.G., Radusky, L.G., Turjanski, A., Wolynes, P.G., Ferreiro, D.U., (2011). Protein frustratometer: a tool to localize energetic frustration in protein molecules. *Nucleic Acids Res.*, **40**, W348–351.

Paper 2: Experimental characterization of the interaction between the N-terminal SH3 domain of CrkL and C3G

International Journal of *Molecular Sciences*

MDPI

Article

Experimental Characterization of the Interaction between the N-Terminal SH3 Domain of Crkl and C3G

Livia Pagano, Francesca Malagrinò, Caterina Nardella, Stefano Gianni * and Angelo Toto *

Istituto Pasteur—Fondazione Cenci Bolognetti, Dipartimento di Scienze Biochimiche "A. Rossi Fanelli" and Istituto di Biologia e Patologia Molecolari del CNR, Sapienza Università di Roma, 00185 Rome, Italy; livia.pagano@uniroma1.it (L.P.); francesca.malagrino@uniroma1.it (F.M.); caterina.nardella@uniroma1.it (C.N.)
* Correspondence: stefano.gianni@uniroma1.it (S.G.); angelo.toto@uniroma1.it (A.T.)

Abstract: Crkl is a protein involved in the onset of several cancer pathologies that exerts its function only through its protein–protein interaction domains, a SH2 domain and two SH3 domains. SH3 domains are small protein interaction modules that mediate the binding and recognition of proline-rich sequences. One of the main physiological interactors of Crkl is C3G (also known as RAPGEF1), an interaction with key implications in regulating cellular growth and differentiation, cell morphogenesis and adhesion processes. Thus, understanding the interaction between Crkl and C3G is fundamental to gaining information about the molecular determinants of the several cancer pathologies in which these proteins are involved. In this paper, through a combination of fast kinetics at different experimental conditions and site-directed mutagenesis, we characterize the binding reaction between the N-SH3 domain of Crkl and a peptide mimicking a specific portion of C3G. Our results show a clear effect of pH on the stability of the complex, due to the protonation of negatively charged residues in the binding pocket of N-SH3. Our results are discussed under the light of previous work on SH3 domains.

Keywords: kinetics; site-directed mutagenesis; stopped-flow

Citation: Pagano, L.; Malagrinò, F.; Nardella, C.; Gianni, S.; Toto, A. Experimental Characterization of the Interaction between the N-Terminal SH3 Domain of Crkl and C3G. *Int. J. Mol. Sci.* **2021**, *22*, 13174. https://doi.org/10.3390/ijms222413174

Academic Editor: Yuri V. Sergeev

Received: 27 October 2021
Accepted: 3 December 2021
Published: 7 December 2021

1. Introduction

Crkl is a ubiquitously expressed 39 kDa adapter protein, member of the proto-oncogene CRK family, that mediates and regulates the linking of several signaling proteins. It was originally discovered in cells from patients with chronic myelogenous leukemia, and its overexpression is correlated with the onset of a number of cancer diseases (recently reviewed in [1]). Crkl plays an essential role in regulating several physiological pathways linked to cytoskeletal changes and cell migration and possesses a prominent role in the onset of human cancers, such as, for example, chronic myelogenous leukemia [2]. Intriguingly, Crkl possesses no catalytic or transcriptional activity and exerts its functions through its protein–protein interaction modules that compose the entire protein, i.e., one N-terminal SH2 domain followed by two SH3 domains (namely N-SH3 and C-SH3).

SH3 domains are small protein interaction modules composed of a five strands β-sandwich and a 3_{10} helix, that typically mediate the binding and recognition of proline-rich sequences, in particular those characterized by the P-X-X-P consensus (X being any amino acid) [3,4], although atypical binding sequences interacting with SH3 have been identified [5,6]. Importantly, the P-X-X-P motif can be arranged into two opposite orientations, defined by the formation of a salt bridge between a positively and a negatively charged residue(s) on the SH3 binding surface [7]. Crk and Crkl display an overlapping list of cellular interactors (such as for example SOS, EPS15 and C3G) [8], which may be explained by their ability to recognize and bind similar consensus sequences P-X-L-P-X-K (Proline-X-Leucine-Proline-X-Lysine) [9,10].

C3G (also known as RAPGEF1, Rap guanine nucleotide exchange factor 1) is the first guanine nucleotide exchange factor discovered to interact with the SH3 domain of

CRK [11,12] and to activate Rap1 GTPases [12]. Rap1 regulates several physiological pathways in the cell, ranging from growth and differentiation to cell morphogenesis and adhesion processes [13]. C3G is a small GTPase with a key role in cell adhesion and cell–cell junction formation [14]. C3G is characterized by a specific region that catalyzes the exchange reaction and several polyproline regions conforming to the consensus P-X-X-P-X-K, which allows it to be recognized and bound by SH3 domain containing proteins. Crkl interacts with C3G through its N-terminal SH3 domain. The association of the Crkl-C3G complex with proteins characterized by the presence of a phosphorylated tyrosine has been proposed as basis of the phosphorylation of a specific tyrosine of C3G and consequent activation of Rap1 [15,16]. A lysine residue in position +2 to the P-X-X-P motif has been demonstrated to be of key importance for binding specificity with Crk [17,18]. Structural analysis of the complex reported that the positive charge carried by the lysine is coordinated by three negatively charged residues in the binding pocket of Crk. Those acidic residues are conserved in Crkl.

Because of its involvement in the onset of several cancer pathologies and its functions solely based on mediating protein-protein interactions, a deep understanding of the mechanism of binding of Crkl with its ligands is of primary importance as a first step toward the definition of potential therapeutic strategies aimed to modulate those interactions. In this paper we characterize the binding reaction occurring between the N-terminal SH3 domain of Crkl and a peptide mimicking a specific region of C3G, ranging from residue 277–296, namely $C3G_{277–296}$, (sequence VVDNSPPPALPPKKRQSAPS) through a combination of fast kinetic binding experiments conducted at different experimental conditions and site-directed mutagenesis. Our kinetic analysis demonstrates a clear effect of pH on the stability of the complex, allowing us to ascribe this effect on the protonation of negatively charged residues. By removing these negative charges through site-directed mutagenesis, we could characterize their specific role in the binding event. Our results and their implications are then discussed under the light of previous work on SH3 domains.

2. Results and Discussion

2.1. The Recognition Event of C3G by the N-SH3 Domain Is Electrostatically Driven

In order to characterize the binding reaction between the N-SH3 domain of Crkl and C3G, we conducted kinetic binding experiments with a stopped-flow apparatus, by rapidly mixing the N-SH3 domain versus the $C3G_{277–296}$ peptide, the latter covalently linked with a dansyl group at its N-terminus (Dansyl-VVDNSPPPALPPKKRQSAPS). This modification allowed us to monitor the binding reaction by following the change of Förster resonance energy transfer (FRET) signal upon binding, which generates from the two naturally present tryptophan residues in the N-SH3 domain in position 164 and 165 (donor) and the dansyl group linked to the peptide (acceptor).

To characterize the binding between N-SH3 and the $C3G_{277–296}$ peptide, we resorted to carry out pre-steady state rapid mixing stopped-flow experiments. In these types of experiments, a common practice lies in performing the kinetic analysis under the so-called pseudo-first-order conditions, i.e., a condition in which one of the two reactants is held at a much higher concentration than the other. In practice, in many cases, it is extremely difficult to achieve such conditions [19]. This is particularly true in cases in which the observed rate constants are so high to approach the experimental limitations of the stopped-flow apparatus. Accordingly, as described below, the analysis of the binding data must be performed by applying the analytical solution of the bimolecular binding transition [19,20].

We rapidly mixed a constant concentration of $C3G_{277–296}$ peptide (0.5 μM) versus increasing concentrations of N-SH3 (ranging from 0.5 to 5 μM) and the observed rate constants (k_{obs}) obtained at different ionic strength conditions (buffer Hepes 50 mM, pH 7.0, 0.15 M, 0.3 M, 0.5 M and 1 M NaCl) were plotted as function of concentration of N-SH3 and fitted with a linear equation (Figure 1). By following previously derived equations [19,20], the dependence of k_{obs} as function of [N-SH3] was fitted with Equation (1), taking into account the non-pseudo-first order conditions [19], allowing us to calculate

the microscopic association rate constant (k_{on}), the microscopic dissociation rate constant (k_{off}) of the binding reaction and the equilibrium dissociation rate constant K_D, such as k_{off}/k_{on}. To increase the reliability of the calculated K_D, we resorted to directly obtaining a k_{off} value through displacement experiments (Table 1), in which a preincubated complex of N-SH3 domain and dansylated $C3G_{277-296}$ peptide (at the concentration of 0.5 and 2 μM, respectively) were rapidly mixed versus a high excess of nondansylated peptide (ranging from 20 to 40 μM). In agreement with the theory in [21], the observed rate constants were insensitive to displacer concentration. In all binding and displacement experiments conducted, traces were satisfactorily fitted with a single-exponential equation.

$$k_{obs} = \left({k_{on}}^2 * (2 - [NSH3])^2 + {k_{off}}^2 + 2k_{on}k_{off}(2 + [NSH3])\right)^{1/2} \quad (1)$$

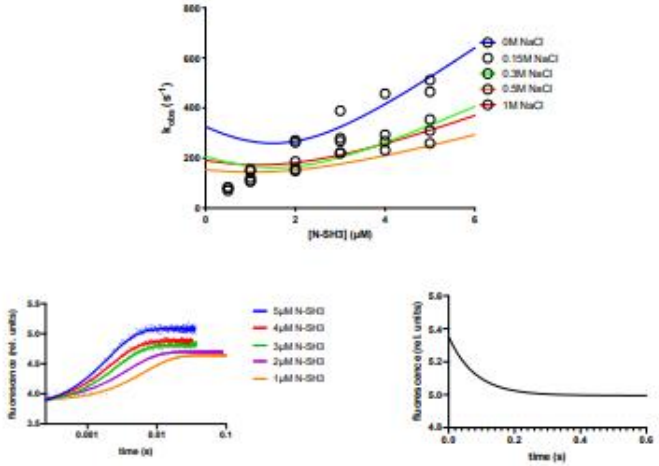


Figure 1. (**Top**) Kinetic binding experiments performed by mixing a constant concentration of dansylated $C3G_{277-296}$ peptide (0.5 μM) versus increasing concentrations of N-SH3 in buffers containing different NaCl concentrations (see legend). Lines are the best fit to Equation (1). (**Bottom left**) Average kinetic traces obtained in binding experiments between dansylated $C3G_{277-296}$ peptide and different concentration of N-SH3 at different concentrations in buffer Hepes 50 mM pH 7.0. Lines are the best fit to a single-exponential equation. (**Bottom right**) Average displacement kinetic trace obtained mixing a preincubated complex of N-SH3 domain and dansylated $C3G_{277-296}$ peptide versus a high excess of nondansylated $C3G_{277-296}$ peptide in buffer Hepes 50 mM pH 7.0. Line represents the best fit to a single-exponential equation.

It should be noticed that the observed rate constants reported in Figure 1 are at the limit of the experimental detection by stopped-flow experiments. Consequently, all the binding experiments reported in this work were measured at 10 °C, in order to slow down the apparent transitions. Hence, it is important to note that the experimental conditions significantly deviate from the physiological conditions, and their interpretation should be mainly taken with comparative purposes.

Changing the ionic strength of the solution is the simplest way to modulate an electrostatically driven binding reaction. It is generally known that shielding the electrostatic attraction of diffusion controlled binding reactions leads to the opposite effects on the microscopic association and dissociation rate constants, with k_{on} rapidly decreasing upon increasing ionic strength [22]. In in vitro experiments, the interaction between two proteins is the result of a random collision forming an encounter complex, stabilized in a final com-

plex after one (or more) transition state(s). In such scenarios, the early events of recognition between charged residues of the two proteins are diffusion controlled and affected by ions dissolved in solution. Then, the binding follows, with desolvation of polar and charged residues at the interface between the two proteins, the bound complex remaining mostly unaffected by increasing concentration of salt in solution. Accordingly, in the case of N-SH3 domain binding with $C3G_{277–296}$, the inspection of Figure 1 and the analysis of kinetic data at different ionic strength conditions (reported in Table 1) display a pronounced effect of salt concentration on the k_{on}.

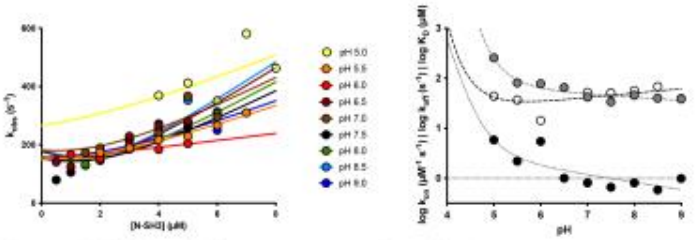


Figure 2. (**Left**) Kinetic binding experiments performed by mixing a constant concentration of dansylated $C3G_{277–296}$ peptide (0.5 μM) versus increasing concentrations of N-SH3 at different pH conditions (see legend). Lines are the best fit to Equation (1). (**Right**) Dependence of logarithm of k_{on} (white) and k_{off} (gray) and equilibrium dissociation rate constant (black) as a function of pH. Lines are the best fit to Henderson–Hasselbalch equation.

2.2. Protonation of Negatively Charged Residues Abolishes Binding

To further investigate the role of charges in the binding reaction between N-SH3 domain and C3G, we resorted to performing kinetic binding experiment at different pH conditions, at a range of pH between 5.0 and 9.0. The observed rate constants obtained at different pH conditions were fitted with Equation (1) (Figure 2, left), and the calculated kinetic parameters k_{on} and k_{off} (obtained from separated displacement experiments) are reported in Table 1.

The dependence of k_{on} and k_{off} as function of pH, reported in Figure 2, right, clearly shows that the affinity between the two interacting molecules decreases with decreasing pH. Unfortunately, the very high value of k_{off} prevents any reliable analysis using a Henderson–Hasselbalch equation. In fact, in the case of the N-SH3, we could not obtain a sigmoidal profile, due to the impossibility to measure binding at pH < 5.0. Although the beginning of a transition is clearly visible, an accurate fit would require more data points for pH < 5.0. On the basis of the available data, it may be concluded that acidic pH conditions cause a dramatic increase in the k_{off}, with the k_{on} being less affected, determining a pronounced destabilization of the complex.

The analysis of these data provides us with important information about the binding mechanism of the N-SH3 domain with C3G. In light of what was previously shown for the ionic strength dependence, kinetic data obtained at different pH highlight a double role for salt bridges formation in both the early recognition events and the stabilization of the complex. Moreover, since SH3 domains have evolved to recognize and bind polyproline sequences, and proline binding occurs mainly through C-H·π interactions with aromatic residues [23,24], the formation of canonical salt bridges may give a substantial contribution in optimizing the recognition and binding of the substrate, improving the specificity in the crowded intracellular environment. Interestingly, the comparison of the primary structures of Crk and Crkl highlights the conservation of three negatively charged residues (D147, E149 and D150 on Crk, D138, E140 and D141 on Crkl) physically located at the binding interface of the proteins (Figure 3). Based on this evidence and on previous structural work on Crk [18], all together our results suggest that D138, E140 and D141 residues of the

N-SH3 domain of Crkl may be responsible for a salt bridge formation with a positively charged residue on C3G.

Table 1. Kinetic parameters at different ionic strength conditions and different pH (in presence of 0.5 M NaCl) obtained from linear fitting of data reported in Figures 1 and 2. The NaCl concentrations reported in the left column were added to buffer Hepes 50 mM pH 7.0.

N-SH3 WT versus $C3G_{277–296}$			
[NaCl]	k_{on} ($\mu M^{-1} s^{-1}$)	k_{off} (s^{-1})	K_D (μM)
0 M	130 ± 7	12 ± 1	0.09 ± 0.05
0.15 M	96 ± 8	23 ± 1	0.24 ± 0.06
0.3 M	83 ± 7	32 ± 2	0.39 ± 0.14
0.5 M	51 ± 4	34 ± 1	0.66 ± 0.12
1.0 M	68 ± 5	51 ± 2	0.75 ± 0.20
pH	k_{on} ($\mu M^{-1} s^{-1}$)	k_{off} (s^{-1})	K_D (μM)
5.0	44 ± 15	257 ± 10	5.9 ± 0.8
5.5	37 ± 3	81 ± 3	2.2 ± 1.0
6.0	15 ± 4	78 ± 3	5.4 ± 1.5
6.5	65 ± 4	65 ± 2	1.0 ± 0.6
7.0	52 ± 5	42 ± 2	0.8 ± 0.5
7.5	51 ± 5	34 ± 2	0.7 ± 0.3
8.0	56 ± 5	46 ± 3	0.8 ± 0.4
8.5	68 ± 7	40 ± 3	0.6 ± 0.4
9.0	39 ± 4	39 ± 2	1.0 ± 0.4

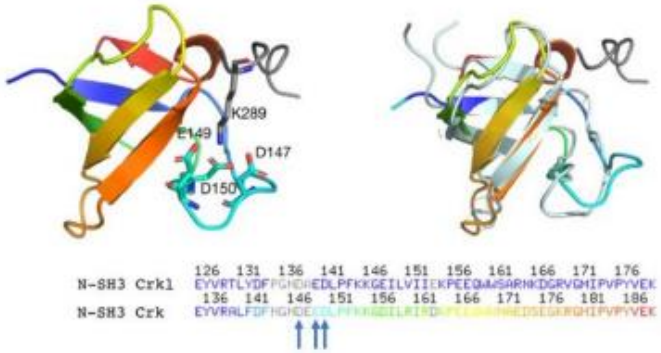


Figure 3. Left: three-dimensional structure of the N-SH3 domain of Crk (rainbow colors) in complex with C3G (gray) (PDB: 1cka). Acidic residues D147, E149 and D150 together with basic residue K289 of C3G are highlighted in sticks. K289 is coordinated by the three negatively charged residues in the binding pocket of Crk, which are conserved in Crkl; **right and bottom**: structural and sequence alignment of the N-SH3 domain of Crk (rainbow colors) with the N-SH3 domain of Crkl (light blue). D147, E149 and D150 residues of Crk are highlighted with blue arrows.

2.3. Determining the Role of D138, E140 and D141 by Site-Directed Mutagenesis

In an effort to analyze the mechanistic role of the residues D138, E140 and D141 in the binding reaction with C3G, we performed site-directed mutagenesis and generated

the D138A, E140A and D141A variants of the N-SH3 domain. At first, to monitor the effect of these mutations on the stability of the domain, we performed (un)folding kinetic experiments in buffer Hepes 50 mM pH 7.0 at 25 °C. The dependence of the logarithm of the observed rate constants (k_{obs}) obtained at different [GdnHCl] (chevron plot) for the wild-type, and the three variants are reported in Figure 4. All the chevron plots were globally fitted by sharing kinetic m-values [25] with Equation (2) describing a two-state scenario, suggesting the absence of intermediate(s) populating along the reaction pathway [26,27]. Importantly, none of the three mutations cause a disruption of the native state, albeit D141A mutation appears mildly destabilizing compared to D138A and E140A.

$$k_{obs} = k_f^0 exp\left(-m_f[GndHCl]/RT\right) + k_u^0 exp(m_u[GndHCl]/RT) \quad (2)$$

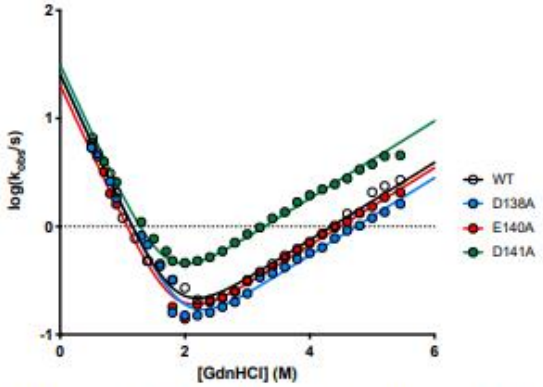


Figure 4. Dependence of the logarithm of the observed rate constants for unfolding and refolding experiments as a function of [GdnHCl] for WT, D138A, E140A and D141A (see legend) obtained in buffer Hepes 50 mM pH 7.5 at 25 °C. Data were globally fitted by sharing kinetic m-values. Lines are the best fit to an equation describing a two-state folding mechanism. WT—k_f = 25.3 ± 2.0 s^{-1}, k_u = 0.027 ± 0.002 s^{-1}; D138A—k_f = 25.9 ± 2.2 s^{-1}, k_u = 0.019 ± 0.001 s^{-1}; E140A—k_f = 20.2 ± 1.6 s^{-1}, k_u = 0.024 ± 0.001 s^{-1}; D141A—k_f = 31.2 ± 2.7 s^{-1}, k_u = 0.066 ± 0.004 s^{-1}; globally shared m-values—m_f = 1.68 ± 0.04 kcal mol^{-1} M^{-1}, m_u = 0.49 ± 0.01 kcal mol^{-1} M^{-1}.

Then, we employed D138A, E140A and D141A variants in kinetic binding experiments with C3G, and we explored the effect of increasing ionic strength on the binding reaction of these variants. The results obtained are reported in Figure 5, and the calculated kinetic data are listed in Table 2. The inspection of Figure 5 and analysis of kinetic data highlight the D138A variant to be affected by increasing salt concentrations, while E140A shows no evident effects. A comparison of kinetic data obtained in the absence of NaCl shows that whilst the microscopic association rate constants calculated for D138A and E140A variants is comparable with the one obtained for the wt, an increase in k_{off} is appreciable. All these aspects suggest a key role of E140 in the recognition event, with D138A being involved mainly in the late events. This scenario is further confirmed by the evidence of a rapidly decreasing k_{on} upon increasing ionic strength of the solution for D138A variant, with the electrostatic charges carried by D138 residue not being involved in the recognition of a positively charged residue on C3G. On the other hand, the overall absence of effect of salt dependence on binding kinetics upon removal of E140 negative charge suggests the interactions formed by this residue acting as a prominent determinant of the early events of the binding reaction between the N-SH3 domain and C3G.

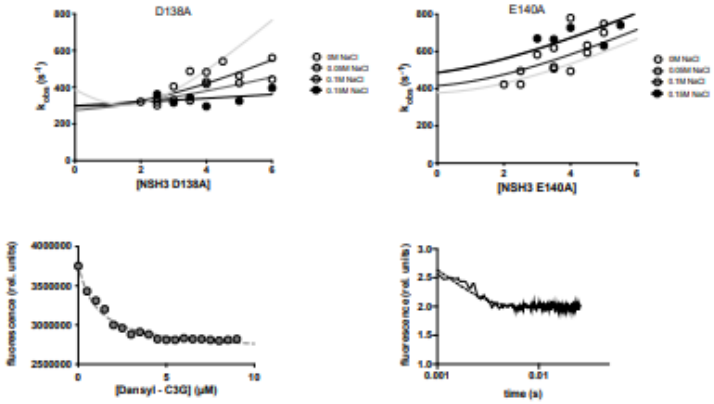


Figure 5. (**Top left**) Kinetic binding experiments performed by mixing a constant concentration of dansylated $C3G_{277–296}$ peptide (0.5 μM) versus increasing concentrations of N-SH3 D138A at different ionic strength conditions (see legend). Lines are the best fit to Equation (1). (**Top right**) Kinetic binding experiments performed by mixing a constant concentration of dansylated $C3G_{277–296}$ peptide (0.5 μM) versus increasing concentrations of N-SH3 E140A at different ionic strength conditions (see legend). Lines are the best fit to Equation (1). (**Bottom left**) Equilibrium binding experiment between N-SH3 D141A held at constant concentration and increasing concentrations of dansylated $C3G_{277–296}$ peptide. Line is the best fit to a hyperbolic function. (**Bottom right**) Average displacement trace obtained for D141A. See text for details of conditions used. Lines are the best fit to a single-exponential function.

Table 2. Kinetic parameters obtained from linear fitting of data reported in Figure 5 (Top left, Top right). Equilibrium dissociation rate constant K_D were calculated as k_{off}/k_{on}.

N-SH3 D138A versus $C3G_{277–296}$			
[NaCl]	**k_{on} ($\mu M^{-1} s^{-1}$)**	**k_{off} (s^{-1})**	**K_D (μM)**
0 M	158 ± 12	139 ± 1	0.9 ± 0.3
0.05 M	81 ± 8	212 ± 2	2.6 ± 0.6
0.10 M	48 ± 8	183 ± 2	3.8 ± 0.3
0.15 M	12 ± 5	254 ± 2	22.0 ± 2.0
N-SH3 E140A versus $C3G_{277–296}$			
[NaCl]	**k_{on} ($\mu M^{-1} s^{-1}$)**	**k_{off} (s^{-1})**	**K_D (μM)**
0 M	78 ± 6	254 ± 3	3.3 ± 0.7
0.05 M	89 ± 8	330 ± 3	3.7 ± 0.6
0.10 M	77 ± 8	302 ± 2	3.9 ± 0.8
0.15 M	81 ± 8	339 ± 2	4.2 ± 1.0

It is of particular interest to discuss the effect of D141A mutation. When employed in kinetic binding experiments, this variant did not return any measurable trace describing a change in FRET signal upon mixing, possibly due to an overall destabilization of the complex and/or sub-millisecond kinetics that could not be resolved by the stopped-flow. To further investigate this aspect and obtain a quantitative measure of the binding affinity of D141A variant with C3G, we resorted to conduct equilibrium binding experiments. Experiments were conducted by exciting samples containing a constant concentration (1 μM) of N-SH3 D141A at 280 nm and following the progressive quenching of two tryptophan residues (in position 164 and 165) fluorescence emission at increasing concentrations of

dansyl-C3G$_{277-296}$. The dependence of fluorescence measured at 350 nm as function of dansyl-C3G concentration is reported in Figure 5. Interestingly, the fitting of data with a hyperbolic equation returned a $K_D = 1.07 \pm 0.05$ μM, demonstrating that D141A variant is capable of binding. On the other hand, although this value reflects a relatively high affinity in the low μM range, it appears to be 10-fold higher than what was measured about wt from kinetic data ($k_{off}/k_{on} = K_D = 0.09 \pm 0.05$ μM). Such a decrease in binding affinity may be at the basis of our impossibility to time-resolve the binding reaction with a stopped-flow apparatus, with the reaction possibly occurring in the dead time of the instrument.

To further investigate this aspect, we resorted to performing displacement experiments, targeted to the direct calculation of microscopic dissociation rate constant k_{off}. We challenged a preincubated complex of D141A NSH3 domain (2 μM) and dansyl-C3G$_{277-296}$ (10 μM) versus a high excess of nondansylated C3G (50 μM), and we could not measure any change in fluorescence emission, suggesting the k_{off} being too high to be measured at the stopped-flow. To test this and to slow down the diffusion of molecules, we repeated the experiment increasing the viscosity of the solution adding 20% w/v sucrose to the buffer Hepes 50 mM pH 7.0. As expected, the increase in solution viscosity allowed us to obtain a displacement trace (shown in Figure 5 bottom right panel). Although the trace was satisfactorily fitted with a single-exponential equation, the k_{off} calculated is 800 ± 30 s^{-1}, a value that is far beyond the resolution capability of the stopped-flow, a consistent part of the reaction occurring in the dead time of the instrument. Based on the calculated affinity of D141A variant for C3G in the absence of sucrose, this dramatic increase in k_{off} might be accompanied by a strong increase in k_{on}.

It is of particular interest to compare the effect of the three site-directed mutants described in this work. In fact, whilst a specific role in the early and late events of the binding reaction could be determined for D138 and E140 residues, in the case of D141, both microscopic association and dissociation rate constants were affected upon mutation. Hence, whereas D138 plays a key role in the stabilization of the formed complex and E140 is mainly involved in the early recognition events of the binding reaction, the D141 sidechain appears to play a key role in both events, in particular in the formation of electrostatic interactions occurring downhill the main energy barrier of the reaction. The increase in k_{on} which must occur in order to maintain a relatively high affinity suggests that the absence of the negative charge in position 141 strongly improves the early recognition event of C3G. Overall, our data support a scenario in which, aside from the nonspecific ionic interactions occurring between N-SH3 and C3G, an additional specific step occurs, driven at least in part by D141 residue, contributing to the final stabilization of the formed complex.

2.4. Determinants of N-SH3 Domain Binding Selectivity

SH3 domains are widespread protein–protein interaction domains. Their main biochemical property relies in the recognition of proline rich sequences, generally identified with the P-X-X-P consensus. Given their fundamental importance in many physiological and molecular pathways in the eucaryotic cell and their role in several human pathologies, SH3 domain gained a strong attention of the scientific community since their discovery, and many of them have been characterized in their mechanisms of interaction with their ligands [4,28–32]. The general structural properties of the recognition of ligands by SH3 domains are well established [30]. However, understanding the molecular determinants of specificity of SH3 domains in general is a difficult task to address, given the large amount of atypical consensus sequences that have been discovered [28].

Our group previously described in detail the mechanism of interaction of the C-SH3 domain of Grb2 with Gab2 [31,32], which is regulated by a complex allosteric mechanism. An analysis of the structure of C-SH3:Gab2 complex (PDB: 2vwf) highlights the presence of negatively charged residues in the binding pocket in direct contact with basic residues of Gab2. Importantly, the topological distribution of these negative charges appears conserved between the C-SH3 domain of Grb2 and the N-SH3 domain of Crk. Ionic strength and pH dependence analysis of the binding reaction, together with mutational analysis presented

in this paper, highlight a prominent role of D138, E140 and D141 residue of Crkl in the binding of $C3G_{277-296}$. Our data show E140 being mainly involved in the early events of recognition of a positively charged residue carried by C3G and D138 with a stabilizing effect on the formed complex. Although we could not resolve kinetics for D141A variant, the data obtained from equilibrium binding experiments and displacement experiments show an effect on binding affinity. In analogy to what was previously found for the N-SH3 domain of Crk [18] and in light of the structural and sequence alignment (Figure 3) our data suggest that D138, E140 and D141 residues may coordinate K289 residue of C3G through salt bridges formation. This electrostatic attraction driving this fundamental protein–protein interaction may represent a key aspect of a conserved mechanism of binding in the Crk family, although on the other hand, it raises questions about how promiscuity is avoided in the intracellular milieu. Future work based on structural determination of the N-SH3:C3G complex and on extensive site-directed mutagenesis would allow us to characterize the selectivity determinants of the N-SH3 domain of Crkl, determine which specific positive residue on C3G plays a role in the recognition, and pinpoint possible long-range allosteric regulation of the binding (described also for other small protein–protein interaction modules, such as PDZ domains [33,34]) occurring simultaneously and/or finely tuning the binding interface of the domain.

3. Conclusions

Achieving a deep understanding of the interaction occurring between Crkl protein and its ligands is of fundamental importance to gaining useful information about the molecular basis of several physiological pathways and human pathologies in which this protein is involved. The employment of rigorous kinetic characterization of the binding reaction at different experimental conditions, together with site-directed mutagenesis, allowed us to describe in detail the roles of electrostatic forces occurring between the N-SH3 domain and a peptide mimicking one of its physiological partners, C3G. Importantly, whilst the SH3 domains are generally thought to recognize their ligands via the P-X-X-P recognition motif, our data exemplify the existence of a negatively charged stretch in the SH3, which is critical in determining the affinity between the interacting molecules. In this view, our study complements and enriches the structural knowledge on this important protein system, by providing a mechanistic insight on the role of these residues, as probed by the effect of site-directed mutagenesis on the association and dissociation rate constant, respectively. Our data represent a first step for future structural and extensive mutational characterization of this protein system.

4. Materials and Methods

4.1. Site-Directed Mutagenesis

The construct encoding the N-SH3 domain of Crkl was subcloned in a pET28b+ plasmid vector. The constructs encoding D138A, E140A and D141A were obtained through site-directed mutagenesis using the QuikChange Lightning Site-Directed Mutagenesis kit (Agilent technologies) according to the manufacturer's instructions. All the mutations were confirmed by DNA sequencing.

4.2. Protein Expression and Purification

The expression of all the His-tagged constructs was performed in *E. coli* cells, strain BL21. Bacterial cells were grown in LB medium, with 30 µg/mL of kanamycin, at 37 °C until OD_{600} = 0.7−0.8 and then induced with 0.5 mM IPTG. The cultures were grown at 37 °C for three hours after induction, kept at 25 °C overnight and then collected by centrifugation. Purification was performed resuspending the pellet in 50 mM TrisHCl, 0.3 M NaCl, pH 7.5 buffer with the addition of antiprotease tablet (Complete EDTA-free, Roche), and then sonicated and centrifuged. The soluble fraction from bacterial lysate was loaded onto a nickel-charged His-Trap chelating HP (GE Healthcare) column equilibrated with 50 mM TrisHCl, 0.3 M NaCl and pH 7.5. Protein was then eluted with a gradient from

0 to 0.5 M imidazole by using an ÄKTA-prime system. Fractions containing the protein were collected, and the imidazole was removed using a HiTrap Desalting column (GE Healthcare), with the protein purified in the final buffer of TrisHCl 50 mM, NaCl 0.3 M, pH 7.5. The purity of the proteins was analyzed through SDS-page.

Peptides mimicking the portion of C3G ranging from residue 277 to 296 (sequence VVDN5PPPALPPKKRQ5APS) in their dansylated and nondansylated variants were purchased from GenScript Biotech.

4.3. Stopped-Flow (un)folding Experiments

Kinetic (un)folding experiments were performed on an Applied Photophysics Pi-star 180 stopped-flow apparatus, monitoring the change of fluorescence emission, exciting the sample at 280 nm and recording the fluorescence emission by using a 320 nm cutoff glass filter. In all experiments, performed at 25 °C in buffer 50 mM Hepes pH 7.5, refolding and unfolding were initiated by an 11-fold dilution of the denatured or the native protein with the appropriate buffer (0 M and 6 M Guanidine HCl). For each denaturant concentration, at least five individual traces were averaged, and the final protein concentration was 1.5 μM. The fluorescence time courses obtained was satisfactorily fitted by using a single-exponential equation. The chevron plots obtained were fitted using an equation describing a two-state folding mechanism.

4.4. Stopped-Flow Kinetic Binding and Displacement Experiments

Kinetic binding experiments were performed on a single-mixing SX-18 stopped-flow instrument (Applied Photophysics), by mixing a constant concentration (0.5 μM) of C3G dansylated versus increasing concentrations of N-SH3 at 10 °C. Ionic strength dependence of wt N-SH3 was performed with concentrations ranging from 1 to 5 μM for 0 M NaCl condition and from 0.5 to 5 μM for 0.15 M, 0.3 M, 0.5 M and 1 M NaCl, buffer Hepes 50 mM pH 7.0. The excitation wavelength was 280 nm, and fluorescence was collected using a 455 nm cut-off filter. At least, five independent acquisitions were collected and averaged for each experiment. The resulting averages were all satisfactorily fitted with a single-exponential equation. Buffers used for pH dependence were: 50 mM Sodium Acetate pH 5.0, 50 mM Sodium Acetate, pH 5.5, Bis–Tris 50 mM pH 6.0, Bis–Tris 50 mM pH 6.5, Hepes 50 mM pH 7.0, Hepes 50 mM pH 7.5, Tris–HCl 50 mM pH 8.0, Tris–HCl 50 mM pH 8.5, Tris–HCl 50 mM pH 9.0, all with added 0.5 M NaCl. Concentrations of N-SH3 used for each experiment were: 4.0–8.0 μM at pH 5.0, 2.0–7.0 μM at pH 5.5, 1.0–5.0 μM at pH 6.0, 0.5–5.0 μM at pH 6.5, 1.5–6.0 μM at pH 7.0, 0.5–5.0 μM at pH 7.5, 1.5–6.0 μM at pH 8.0, 1.5–6.0 μM pH 8.5, 1.5–6.0 μM at pH 9.0. For the ionic strength dependence of D138A and E140A variants buffer used were 0.05 M NaCl, 0.1 M NaCl, 0.15 M NaCl, in 50 mM Hepes pH 7.5. Concentrations of N-SH3 D138A used for each experiment were: 2.0–4.5 μM at 0 M NaCl, 2.5–6.0 μM at 0.05 M, 0.1 M, 0.15 M NaCl. Concentrations of N-SH3 E140A used for each experiment were: 2.0–4.5 μM at 0 M NaCl, 3.0–5.5 μM at 0.05 M NaCl, 3.0–5.0 μM at 0.1 M NaCl, 3.0–5.5 μM at 0.15 M NaCl. Displacement kinetic experiments were performed by mixing a preincubated complex of N-SH3 (in all its variants) and dansylated C3G versus a high excess of nondansylated C3G (see text for details). Experiments were performed in the same buffer conditions as the binding experiments, except for D141A variant, for which 20% w/v sucrose was added to the buffer Hepes 50 mM pH 7.0. Displacement traces were fitted with a single-exponential equation.

4.5. Equilibrium Binding Experiment of D141A

Equilibrium experiment on D141A (fixed concentration at 1 μM) was carried out on a Fluoromax single-photon counting spectrofluorometer (Jobin-Yvon, Newark, NJ, USA), by mixing the construct with increasing dansyl-C3G concentrations. Experiments were performed at 10 °C, using a quartz cuvette with a path length of 1 cm, in 50 mM Hepes pH 7.0 and measuring the change in fluorescence of the naturally present tryptophan

residues in position 164 and 165 at increasing concentration of dansyl-C3G. The excitation wavelength was 280 nm, and fluorescence spectra were recorded between 300 and 400 nm.

Author Contributions: Conceptualization, S.G. and A.T.; methodology L.P., F.M., C.N. and A.T.; formal analysis, A.T.; investigation, L.P., F.M. and C.N.; resources, S.G.; writing—original draft preparation, A.T.; writing—review and editing, S.G. and A.T.; funding acquisition, S.G. All authors have read and agreed to the published version of the manuscript.

Funding: Work partly supported by grants from the Italian Ministero dell'Istruzione dell'Università e della Ricerca (Progetto di Interesse 'Invecchiamento' to S.G.), Sapienza University of Rome (RP11715C34AEAC9B and RM1181641C2C24B9, RM11916B414C897E, RG12017297FA7223 to S.G, AR22117A3CED340A to C.N.), by an ACIP grant (ACIP 485-21) from Institut Pasteur Paris to S.G., the Associazione Italiana per la Ricerca sul Cancro (Individual Grant—IG 24551 to S.G.), the Regione Lazio (Progetti Gruppi di Ricerca LazioInnova A0375-2020-36559 to S.G.) the Istituto Pasteur Italia (Teresa Ariaudo Research Project 2018, to A.T.). F.M. was supported by a fellowship from the FIRC—Associazione Italiana per la Ricerca sul Cancro (Filomena Todini fellowship).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Park, T. Crk and CrkL as Therapeutic Targets for Cancer Treatment. *Cells* **2021**, *10*, 739. [CrossRef]
2. Nichols, G.L.; Raines, M.A.; Vera, J.C.; Lacomis, L.; Tempst, P.; Golde, D.W. Identification of CRKL as the Constitutively Phosphorylated 39-KD Tyrosine Phosphoprotein in Chronic Myelogenous Leukemia Cells. *Blood* **1994**, *84*, 2912–2918. [CrossRef] [PubMed]
3. Ren, R.; Mayer, B.J.; Cicchetti, P.; Baltimore, D. Identification of a Ten-Amino Acid Proline-Rich SH3 Binding Site. *Science* **1993**, *259*, 1157–1161. [CrossRef] [PubMed]
4. Yu, H.; Chen, J.K.; Feng, S.; Dalgarno, D.C.; Brauer, A.W.; Schreiber, S.L. Structural Basis for the Binding of Proline-Rich Peptides to SH3 Domains. *Cell* **1994**, *76*, 933–945. [CrossRef]
5. Barnett, P.; Bottger, G.; Klein, A.T.; Tabak, H.F.; Distel, B. The Peroxisomal Membrane Protein Pex13p Shows a Novel Mode of SH3 Interaction. *EMBO J.* **2000**, *19*, 6382–6391. [CrossRef] [PubMed]
6. Nishida, M.; Nagata, K.; Hachimori, Y.; Horiuchi, M.; Ogura, K.; Mandiyan, V.; Schlessinger, J.; Inagaki, F. Novel Recognition Mode between Vav and Grb2 SH3 Domains. *EMBO J.* **2001**, *20*, 2995–3007. [CrossRef]
7. Lim, W.A.; Richards, F.M.; Fox, R.O. Structural Determinants of Peptide-Binding Orientation and of Sequence Specificity in SH3 Domains. *Nature* **1994**, *372*, 375–379. [CrossRef]
8. Uemura, N.; Griffin, J.D. The Adapter Protein Crkl Links Cbl to C3G after Integrin Ligation and Enhances Cell Migration. *J. Biol. Chem.* **1999**, *274*, 37525–37532. [CrossRef] [PubMed]
9. Sparks, A.B.; Rider, J.E.; Hoffman, N.G.; Fowlkes, D.M.; Quillam, L.A.; Kay, B.K. Distinct Ligand Preferences of Src Homology 3 Domains from Src, Yes, Abl, Cortactin, P53bp2, PLCgamma, Crk, and Grb2. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 1540–1544. [CrossRef] [PubMed]
10. Matsuda, M.; Ota, S.; Tanimura, R.; Nakamura, H.; Matuoka, K.; Takenawa, T.; Nagashima, K.; Kurata, T. Interaction between the Amino-Terminal SH3 Domain of CRK and Its Natural Target Proteins. *J. Biol. Chem.* **1996**, *271*, 14468–14472. [CrossRef] [PubMed]
11. Knudsen, B.S.; Feller, S.M.; Hanafusa, H. Four Proline-Rich Sequences of the Guanine-Nucleotide Exchange Factor C3G Bind with Unique Specificity to the First Src Homology 3 Domain of Crk. *J. Biol. Chem.* **1994**, *269*, 32781–32787. [CrossRef]
12. Gotoh, T.; Hattori, S.; Nakamura, S.; Kitayama, H.; Noda, M.; Takai, Y.; Kaibuchi, K.; Matsui, H.; Hatase, O.; Takahashi, H. Identification of Rap1 as a Target for the Crk SH3 Domain-Binding Guanine Nucleotide-Releasing Factor C3G. *Mol. Cell Biol.* **1995**, *15*, 6746–6753. [CrossRef]
13. Bos, J.L.; de Rooij, J.; Reedquist, K.A. Rap1 Signalling: Adhering to New Models. *Nat. Rev. Mol. Cell. Biol.* **2001**, *2*, 369–377. [CrossRef] [PubMed]
14. Raaijmakers, J.H.; Bos, J.L. Specificity in Ras and Rap Signaling. *J. Biol. Chem.* **2009**, *284*, 10995–10999. [CrossRef] [PubMed]
15. Ichiba, T.; Hashimoto, Y.; Nakaya, M.; Kuraishi, Y.; Tanaka, S.; Kurata, T.; Mochizuki, N.; Matsuda, M. Activation of C3G Guanine Nucleotide Exchange Factor for Rap1 by Phosphorylation of Tyrosine 504. *J. Biol. Chem.* **1999**, *274*, 14376–14381. [CrossRef]
16. Sakkab, D.; Lewitzky, M.; Posern, G.; Schaeper, U.; Sachs, M.; Birchmeier, W.; Feller, S.M. Signaling of Hepatocyte Growth Factor/Scatter Factor (HGF) to the Small GTPase Rap1 via the Large Docking Protein Gab1 and the Adapter Protein CRKL. *J. Biol. Chem.* **2000**, *275*, 10772–10778. [CrossRef]
17. Knudsen, B.S.; Zheng, J.; Feller, S.M.; Mayer, J.P.; Burrell, S.K.; Cowburn, D.; Hanafusa, H. Affinity and Specificity Requirements for the First Src Homology 3 Domain of the Crk Proteins. *EMBO J.* **1995**, *14*, 2191–2198. [CrossRef]
18. Wu, X.; Knudsen, B.; Feller, S.M.; Zheng, J.; Sali, A.; Cowburn, D.; Hanafusa, H.; Kuriyan, J. Structural Basis for the Specific Interaction of Lysine-Containing Proline-Rich Peptides with the N-Terminal SH3 Domain of c-Crk. *Structure* **1995**, *3*, 215–226. [CrossRef]

19. Malatesta, F. The Study of Bimolecular Reactions under Non-Pseudo-First Order Conditions. *Biophys. Chem.* **2005**, *116*, 251–256. [CrossRef] [PubMed]
20. Gianni, S.; Engström, A.; Larsson, M.; Calosci, N.; Malatesta, F.; Eklund, L.; Ngang, C.C.; Travaglini-Allocatelli, C.; Jemth, P. The Kinetics of PDZ Domain-Ligand Interactions and Implications for the Binding Mechanism. *J. Biol. Chem.* **2005**, *280*, 34805–34812. [CrossRef] [PubMed]
21. Antonini, E.; Brunori, M. *Hemoglobin and Myoglobin in Their Reactions with Ligands*; North-Holland: Amsterdam, The Netherlands, 1971.
22. Schreiber, G.; Haran, G.; Zhou, H.-X. Fundamental Aspects of Protein−Protein Association Kinetics. *Chem. Rev.* **2009**, *109*, 839–860. [CrossRef]
23. Pal, D.; Chakrabarti, P. Cis Peptide Bonds in Proteins: Residues Involved, Their Conformations, Interactions and Locations. *J. Mol. Biol.* **1999**, *294*, 271–288. [CrossRef]
24. Brandl, M.; Weiss, M.S.; Jabs, A.; Sühnel, J.; Hilgenfeld, R. C-H...Pi-Interactions in Proteins. *J. Mol. Biol.* **2001**, *307*, 357–377. [CrossRef]
25. Myers, J.K. Denaturant m Values and Heat Capacity Changes: Relation to Changes in Accessible Surface Areas of Protein Unfolding. *Protein Sci.* **1995**, *4*, 2138–2148. [CrossRef]
26. Jackson, S.E.; Fersht, A.R. Folding of Chymotrypsin Inhibitor 2. 1. Evidence for a Two-State Transition. *Biochemistry* **1991**, *30*, 10428–10435. [CrossRef]
27. Fersht, A.R.; Matouschek, A.; Serrano, L. The Folding of an Enzyme. I. Theory of Protein Engineering Analysis of Stability and Pathway of Protein Folding. *J. Mol. Biol.* **1992**, *224*, 771–782. [CrossRef]
28. Cesareni, G.; Panni, S.; Nardelli, G.; Castagnoli, L. Can We Infer Peptide Recognition Specificity Mediated by SH3 Domains? *FEBS Lett.* **2002**, *513*, 38–44. [CrossRef]
29. Kay, B.K. SH3 Domains Come of Age. *FEBS Lett.* **2012**, *586*, 2606–2608. [CrossRef]
30. Saksela, K.; Permi, P. SH3 Domain Ligand Binding: What's the Consensus and Where's the Specificity? *FEBS Lett.* **2012**, *586*, 2609–2614. [CrossRef]
31. Malagrinò, F.; Troilo, F.; Bonetti, D.; Toto, A.; Gianni, S. Mapping the Allosteric Network within a SH3 Domain. *Sci. Rep.* **2019**, *9*, 8279. [CrossRef]
32. Toto, A.; Bonetti, D.; De Simone, A.; Gianni, S. Understanding the Mechanism of Binding between Gab2 and the C Terminal SH3 Domain from Grb2. *Oncotarget* **2017**, *8*, 82344–82351. [CrossRef]
33. Gianni, S.; Haq, S.R.; Montemiglio, L.C.; Jürgens, M.C.; Engström, Å.; Chi, C.N.; Brunori, M.; Jemth, P. Sequence-Specific Long Range Networks in PSD-95/Discs Large/ZO-1 (PDZ) Domains Tune Their Binding Selectivity. *J. Biol. Chem.* **2011**, *286*, 27167–27175. [CrossRef]
34. Chi, C.N.; Elfström, L.; Shi, Y.; Snäll, T.; Engström, A.; Jemth, P. Reassessing a Sparse Energetic Network within a Single Protein Domain. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 4679–4684. [CrossRef]

Paper 3: Double mutant cycles as a tool to address folding, binding, and allostery

International Journal of *Molecular Sciences*

MDPI

Review

Double Mutant Cycles as a Tool to Address Folding, Binding, and Allostery

Livia Pagano [1], Angelo Toto [1], Francesca Malagrinò [1], Lorenzo Visconti [1], Per Jemth [2,*] and Stefano Gianni [1,*]

[1] Istituto Pasteur—Fondazione Cenci Bolognetti, Dipartimento di Scienze Biochimiche 'A. Rossi Fanelli' and Istituto di Biologia e Patologia Molecolari del CNR, Sapienza Università di Roma, 00185 Rome, Italy; livia.pagano@uniroma1.it (L.P.); angelo.toto@uniroma1.it (A.T.); francesca.malagrino@uniroma1.it (F.M.); lorenzo.visconti@uniroma1.it (L.V.)

[2] Department of Medical Biochemistry and Microbiology, Uppsala University, SE-75123 Uppsala, Sweden

* Correspondence: Per.Jemth@imbim.uu.se (P.J.); stefano.gianni@uniroma1.it (S.G.)

Abstract: Quantitative measurement of intramolecular and intermolecular interactions in protein structure is an elusive task, not easy to address experimentally. The phenomenon denoted 'energetic coupling' describes short- and long-range interactions between two residues in a protein system. A powerful method to identify and quantitatively characterize long-range interactions and allosteric networks in proteins or protein–ligand complexes is called double-mutant cycles analysis. In this review we describe the thermodynamic principles and basic equations that underlie the double mutant cycle methodology, its fields of application and latest employments, and caveats and pitfalls that the experimentalists must consider. In particular, we show how double mutant cycles can be a powerful tool to investigate allosteric mechanisms in protein binding reactions as well as elusive states in protein folding pathways.

Keywords: coupling energy; site-directed mutagenesis; interaction networks

Citation: Pagano, L.; Toto, A.; Malagrinò, F.; Visconti, L.; Jemth, P.; Gianni, S. Double Mutant Cycles as a Tool to Address Folding, Binding, and Allostery. *Int. J. Mol. Sci.* **2021**, *22*, 828. https://doi.org/10.3390/ijms22020828

Received: 31 December 2020
Accepted: 13 January 2021
Published: 15 January 2021

1. Introduction

A rigorous description of any physical system demands the detailed characterization of its structural elements, as well as the quantitative investigation of its energetic components. In the case of proteins, whilst an arsenal of different experimental techniques has been developed to address the structural features of these molecules [1–3], the balance of forces regulating protein architecture, folding and function is much more difficult to address experimentally. In fact, such an aim is complicated by the very small changes in energy between all the different populated conformational states and their rapid rates of inter-conversion [4]. Furthermore, because protein sequences include a combination of hydrophobic (non-polar) and hydrophilic (polar) regions, the transient interactions of these regions with water molecules result in a very complex scenario and a quantitative prediction of all contributing Gibbs free energies is still elusive.

How can we approach this problem? The so-called 'double mutant cycles' approach is a powerful method to measure the strength of molecular contacts between interacting side-chains [5–9]. This technique is generally based on the synergy between site-directed mutagenesis and quantitative measurement of biophysical properties of a protein system. Here we provide a review focused on this experimental methodology. In particular, we recapitulate the key principles of double mutant cycle analysis, exemplifying some possible applications and we discuss the method in light of published work. Furthermore, we emphasize the key caveats and pitfalls associated with double mutant cycle.

2. Principles of Double Mutant Cycles and Basic Equations

In his seminal analysis of the theory of heredity [10], William Bateson observed that, for several genetic features, the concept of 'dominant' and 'recessive', previously

introduced by Mendel, was inadequate. In fact, in some cases, it was clear that the effect of a mutation in given gene was severely influenced by one or more other genes. In these cases, the term 'epistasis' was suggested, reflecting the condition whereby the phenotypic manifestation of a mutation is dependent on the genetic background in which it appears [11,12]. Analogously, ever since the early days of protein engineering, it was clear that some features of amino-acid side chains in regulating protein function were dependent on other side chains. In fact, only two years after the purification of the very first site-directed mutant, performed on the tyrosyl-tRNA synthetase in 1982 [13], it was observed on the same enzyme that the role of two residues in the active site was affected by the conformational change introduced by a threonine to proline mutation in a distal position, thereby introducing the concept of 'coupling' [14].

It is of critical importance to define quantitatively the extent of coupling between two interacting amino acids as well as to introduce a feasible methodology to measure it, the double mutant cycle [5–9]. In principle, a double mutant cycle assumes that when a perturbation is introduced in two non-interacting residues, X and Y, the effect of each single perturbation is additive in a double mutant where both perturbations are present. Thus, the change in free energy upon mutation of X, associated with any structural or functional behavior of a given protein, may be expressed as $\Delta\Delta G_{P\text{-}XY \rightarrow P\text{-}Y}$. Analogously, the change in free energy associated with the residue Y will be equal to $\Delta\Delta G_{P\text{-}XY \rightarrow P\text{-}X}$. If the mutations are independent of each other, then a double mutant of X and Y will display as:

$$\Delta\Delta G_{P\text{-}XY \rightarrow P} = \Delta\Delta G_{P\text{-}XY \rightarrow P\text{-}Y} + \Delta\Delta G_{P\text{-}XY \rightarrow P\text{-}X} \quad (1)$$

Conversely, a non-zero value of:

$$\Delta\Delta\Delta G_{XY} = \Delta\Delta G_{P\text{-}XY \rightarrow P} - \Delta\Delta G_{P\text{-}XY \rightarrow P\text{-}Y} - \Delta\Delta G_{P\text{-}XY \rightarrow P\text{-}X} \quad (2)$$

would correspond to the free energy of interaction between the X and Y and would therefore correspond to the coupling free energy of X and Y with respect to the probed structural or functional property of the protein. The $\Delta\Delta\Delta G_{XY}$ measures therefore the energetic strength of the interaction between positions X and Y and, when its value is different from zero, it represents the interaction (or coupling) energy between the two residues. A scheme summarizing the basic principles and equations of double mutant cycles is reported in Figure 1.

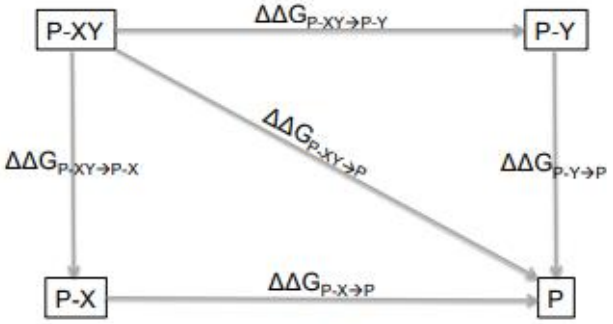


Figure 1. Schematic representation of a double mutant cycle. P-XY is the wild-type protein with the two residues X and Y. P-X and P-Y are the variants where Y and X are mutated, respectively, and P is the corresponding double mutant. The change in free energy upon mutation of X (or Y) is equal to $\Delta\Delta G_{P\text{-}XY \rightarrow P\text{-}Y}$ (or $\Delta\Delta G_{P\text{-}XY \rightarrow P\text{-}X}$) and $\Delta\Delta G_{P\text{-}XY \rightarrow P}$ is the change in free energy of the double mutant.

3. The Double Mutant Cycle: Strengths, Caveats, and Pitfalls

When introducing an experimental technique, it is of particular importance to highlight its major advantages, as well as the caveats and pitfalls. Thus, what is the advantage to perform double mutant cycles as compared to single site mutagenesis? When mutating an amino acid side-chain into another, the change in free energy upon mutation $\Delta G_{N\text{-}N'}$ is in fact the sum of different components [15]: (i) The change in free energy of covalent bonds, ΔG_{cov}; (ii) that arising from the changes in noncovalent bonds located at the site of mutation, ΔG_{noncov}; (iii) any additional free energy changes related to a reorganization of the protein, ΔG_{reorg}; and (iv) changes in solvation energy, ΔG_{solv}. When performing folding and/or binding experiments, the ΔG_{cov} is conserved in all states and, therefore, it cancels out in the analysis. On the other hand, ΔG_{noncov} is generally assumed to be the dominant contributor to $\Delta G_{N\text{-}N'}$. However, the contribution of the four ΔG values may differ for different states. For example, the values of ΔG_{reorg} and ΔG_{solv} may be different in native and denatured states for a folding reaction, such that $\Delta G_{N\text{-}N'}$-$\Delta G_{D\text{-}D'}$ will report not only on the energetics at the mutation site ΔG_{noncov}. This effect will complicate the analysis of the experimental data. A similar situation may occur for bound and free states for a bimolecular interaction.

To exemplify this scenario, let us consider the binding between a protein and a ligand:

$$P_N + L = P_NL \quad (3)$$

When and if a site-directed mutation is introduced in P_N, the folding stability of the protein could be affected, such that its native state is destabilized. Thus, by considering the scheme:

$$P_D = P_N + L = P_NL \quad (4)$$

In the latter case, binding is considered in conjunction with a large conformational change, such as the (un)folding of the protein. In this case, it is clear that the apparent K_D would be affected by the presence of a linked equilibrium, such that:

$$K_D^{app} = K_D \frac{1}{1 + \frac{1}{K_{D-N}}} \quad (5)$$

Thus, if the mutation affects the stability of the native state such that a significant population of the protein is shifted towards the denatured state, the apparent binding constant is also affected; even in cases in which the mutation would not perturb the binding interface at all. In those cases, ΔG_{reorg} and ΔG_{solv} are dominant over the value of ΔG_{noncov}, and any analysis of the local effect of the mutation is prevented.

By considering these premises, it is clear that the double mutant cycle, which relies on calculating the difference between the effects on the double mutant over the two single variants, is extremely powerful. In fact, even in cases in which the mutation results in structural shifts of the protein, as exemplified above, those found in the single mutants are usually present in the double mutant and, therefore, their energetic effects in the double mutant cycle are cancelled out [16–18]. These considerations reinforce the importance of this method as a tool to investigate directly the energetics of interaction between two residues.

Due to the complexity of site directed mutagenesis, it is important to stress that conservative mutations should be used to probe binding or folding [19]. Values of ΔG_{reorg} and ΔG_{solv} for conservative mutations are more likely to be lower than the value of ΔG_{noncov}, which, as briefly discussed above, tends to simplify the analysis. Additionally, since the coupling free energy arises from the difference between the value obtained for the double mutant minus the sum of the values obtained from the single mutants, it might be difficult to calculate a reliable value of $\Delta\Delta G_{XY}$ when the effect of one mutation is very large such that limitations in the biophysical method result in a large experimental error.

For example, isothermal titration calorimetry usually works best in the range 100 nM–1 μM and mutations resulting in a 100-fold destabilization will lead to a large experimental error.

Over and above the theoretical basis of this type of analysis, there are also some practical issues that demand additional consideration. In fact, it is a good advice to choose the mutations properly. Since the double mutant cycle technique is aimed at measuring the free energy of interactions between two residues, it is recommended to keep the two terms ΔG_{reorg} and ΔG_{solv} as low as possible. Consequently, since mutations of large hydrophobic side chains may lead to very large ΔG_{reorg} and ΔG_{solv}, these are not well suited for double mutant cycles. Thus, in analogy to the Φ value analysis in protein folding studies [19], it is recommended to mutate hydrophobic side chains, introducing small side chain deletions and without altering the stereochemistry (i.e., Ile→Val→Ala→Gly; Leu→Ala→Gly; Thr→Ser; Phe→Ala→Gly). Additionally, analogously to what is normally suggested in the case of the Φ value analysis in protein folding, a reliable dataset is generally based on a large number of site-directed mutants, allowing to probe pairwise residue interactions encompassing a relevant fraction of the protein structure.

After the mutations are introduced, it is critical to measure the changes in free energy accurately. This task may be achieved by employing the standard spectroscopic methodologies that are classically used in affinity and/or stability measurements for proteins. These techniques may span from equilibrium titration with spectroscopic techniques such as absorbance, fluorescence, circular dichroism, etc., to other methods, such as calorimetry.

4. Double Mutant Cycles to Understand Intramolecular Interactions

The invention of double mutant cycles was associated with the quantification of intramolecular interactions in proteins. In their pioneering work on protein engineering Winter, Fersht, and co-workers observed that, in the case of a tyrosyl-tRNA synthetase, a variant in which a threonine residue was mutated into proline increased the affinity of the enzyme for ATP [13]. By making mutants of the proline-containing enzyme at two other positions directly involved in ATP binding, it was possible to show that the presence of the proline improves the strength of one of these contacts. On the basis of these observations, it was proposed that the propagation of a structural change in an enzyme induced by mutation could be explored by the use of further mutations, thereby introducing the double mutant cycle methodology as a unique tool for the investigation of intramolecular interactions in proteins.

Following this original work, several protein systems of known structure have been subjected to double mutant cycles [6]. In these cases, the method represents a powerful tool to unveil the details of distinct interactions in protein stability and function. In this context, it is worth to highlight the employment of double mutant cycles to clarify the role of surface and buried salt bridges in protein stability, as well as charge-aromatic interactions. Importantly, the cycles have also been used in more complex approaches with three interacting residues, implying the design of multidimensional cycles. Barnase [20–22], staphylococcal nuclease [23,24], and lambda repressor [25,26] are three examples of proteins subjected to double mutant cycle analysis.

Over and above the cases in which double mutant cycles have been employed to address intramolecular interactions in known structures, it is important to note that, in some context, the approach can be used to investigate elusive states that are difficult to characterize experimentally. The KIX domain is a globular domain that is a part of a large coactivator protein called CBP [27]. The three-dimensional structure of KIX consists of three α-helices and two short 3_{10}– helices (Figure 2). A detailed characterization of its folding pathway revealed that, whilst the protein appears to fold in a simple two-state manner, its denatured state retains a considerable amount of structure [28], as probed by the analysis of the so-called *m* values, a set of parameters related to the change in accessible surface area upon (un)folding [29]. To address the structural details of the residual structure in the denatured state of KIX, double mutant cycles were performed [28]. In particular, by focusing primarily on the single variants that had a major effect on the folding *m* values,

a series of double mutants were produced and characterized. Interestingly, the folding characterization of these mutants allowed mapping the presence of distinct non-native interactions in the denatured state, as well as in measuring their strength in terms of ΔΔΔG. A similar analysis of structure in the denatured state was also performed on the C-terminal domain of nucleophosmin by conducting folding and unfolding experiments on several site-directed variants at different experimental conditions [30].

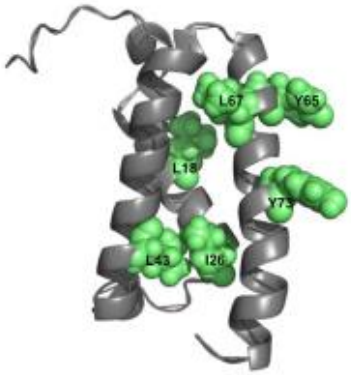


Figure 2. Structural distribution of residues L18, I26, L43, Y65, L67, and Y73 (green spheres) in the KIX domain (PDB 2AGH). The analysis of folding *m* values, in combination with double mutant cycle analysis, allowed to characterize, from a thermodynamic perspective (i.e., measuring ΔΔΔG values), the role of these residues in stabilizing non-native interactions in the denatured state of KIX. With the exception of I26 and L43 that are in direct contact, residues displaying significant coupling free energies are located in different regions of the domain (see [28] for details).

5. Protein Binding and Allostery

Ever since the pioneering work of Kendrew and Perutz, it became clear that a deep understanding of proteins' structure and function would depend on the rigorous description of their dynamic properties. In fact, only the identification of the R and T quaternary states of haemoglobin provided a structural framework to understand its cooperative and allosteric behavior, pinpointing the role of residues 'other' (from the Greek *allo*-) than those in the binding site in regulating function [31]. In this context, when studying the binding of two molecules, it is clear that double mutant cycles may provide a straightforward method to analyze binding and allostery [16,32]. When studying a number of i mutants in the A protein, binding to j mutants of the B protein, it is possible to address i × j double mutant cycles and define their interactions quantitatively. In practical terms, in those cases, it would be required to perform the experiments: (i) on the mutant of protein A versus the wild-type form of the protein B; (ii) on the wild-type of protein A versus the mutant form of the protein B; (iii) on all combinations of the mutant variants of both proteins. If the changes in free energies are not additive, the probed residues are energetically coupled.

The interaction between barnase and barstar represents a paradigm example of the use of double mutant cycles to understand protein-protein binding. In this case, several site-directed mutants were analyzed providing a detailed description of the energetics of such interactions [18,33]. In particular, it was found that the coupling free energy generally depends on the distance between the probed residues. Whilst residues closer than 7 Å tend to interact co-operatively, at greater distances the changes in free energy upon mutations become additive. However, it was found that the interaction network between functional residues is not always trivial [34–37], such that they can be directly deduced from the crystal structure. For example, the salt-bridge between Lys27 in barnase and Asp39 of barstar was found to be very strong, while that between Glu76 of barstar and Arg59 of barnase is

relatively weak, showing the importance of the local environment for binding [38], as well as the need to complement the structural data with biophysical experiments.

The double mutant cycle approach has been subsequently employed on other protein systems, with the specific aim to quantify the interaction networks between interacting protein systems. Relevant examples include coiled-coil XGCN4-p1 [39], as well as characterization of the allosteric network of different PDZ [40–43] and SH3 domains [44]. In all these cases, it was possible to identify energetic coupling between residues that were not necessarily in direct contact in the three-dimensional structure of the respective protein complexes, confirming the concept that allosteric networks tend to follow intricate pathways that demand a careful experimental investigation.

Out of the examples highlighted above, it is particularly interesting to discuss some of the conclusions that have been drawn on PDZ domains [40–43]. In fact, the experimental data suggested that, in the majority of cases, a positive coupling energy for the Val_0 to Abu substitution was observed in both PDZ2 from PTP-BL and PDZ3 from PSD95 (Figure 3), although the absolute values were variable. Thus, the effect of the mutations in the protein on binding was more pronounced when the wild-type peptide was used, as compared to that in which a methyl group was removed from position 0. This intriguing finding suggested that the entire PDZ structure could been under selective pressure to optimize the binding of its physiological ligand. Sequence variation in the protein domain is therefore not neutral to peptide binding, strongly indicating that selectivity by the domain is not solely determined by the subset of residues directly involved in ligand binding. Importantly, this finding appears to support a scenario whereby the cross-talk between binding sites and remote residues may be used to fine tune target selectivity, and possibly to decrease the cross-reactivity between homologous PDZ domains.

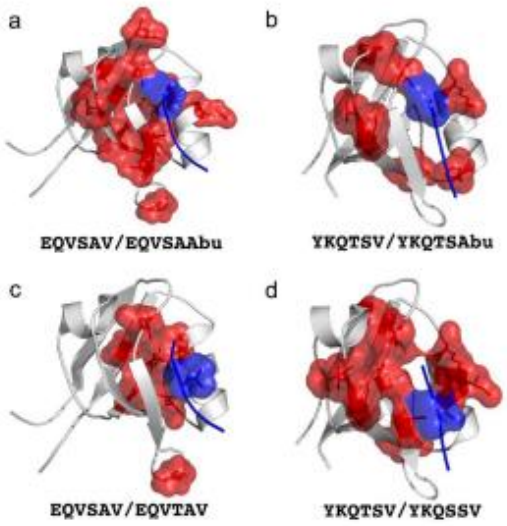


Figure 3. Graphic representation of the different allosteric networks of PDZ domain-ligand complexes of protein tyrosine phosphatase Basophil-like PDZ2 (**a** and **c** panels) and PSD-95 PDZ3 (**b** and **d** panels) with their natural ligand (showed below each panel). All the mutated residues are represented as spheres both in the ligand (blue) and in the PDZ domain (red). The peptide ligand mutations Val→Abu (deletion of a γ-methyl group from Val), Ser→Thr or Thr→Ser were chosen because they interact with the binding pocket of PDZ domains.

The complex nature of the functions of proteins often demands their structure to be composed of multiple domains. In these cases, it is very frequent that function and binding may depend on the interdomain architecture and dynamics, such that the individual domains in isolation display different features compared to when they are present in multidomain constructs. While this concept is widely accepted, our detailed knowledge about the effects of supertertiary structure in folding and function is still limited. In this context, double mutant cycles proved very useful. An extensive double mutant cycle analysis of a PDZ domain, both in isolation and in the context of a supramodule comprising the PDZ domain, an SH3 and a GK domain, showed that allosteric networks are highly sensitive to the supertertiary structure [45]. In particular, it was found that the presence of the SH3–GK tandem resulted in strong coupling from the bound peptide ligand to the β1β2 loop, β2β3 loop, and α3 helix that was not observed with the single PDZ3 domain. These findings prompted the authors to reinforce the importance of extending double mutant cycle experiments to multidomain systems.

6. Double Mutant Cycles In Silico

The double mutant cycle methodology generally relies on the design, production and characterization of protein mutant variants. Nevertheless, an interesting application of this method is an in silico analysis only. Horovitz and co-workers, who investigated the interaction between two cysteine residues, namely Cys137 and Cys518, in the *Escherichia coli* chaperonin GroEL, first introduced this approach [46]. By performing a multiple sequence alignment on proteins belonging to the Hsp60 family, they found that naturally-occurring variations at positions 137 and 518 were likely to be coordinated. An experimental validation of this interaction further substantiated this analysis and suggested that the study of the co-variance of residues at different positions may represent a valuable method to detect interactions in protein systems.

In a seminal study published in Science, Lockless and Ranganathan extended the approach to a larger scale [47]. By using the PDZ domain family as a model system, the authors investigated systematically the evolutionary coupling between the different positions of the domain. The principle of the method is that a statistical coupling between two sites, i and j, may be defined as the degree to which amino acid frequencies at site i change in response to a perturbation of frequencies at another site, j. Then, by focusing on residue His76, which is in an important position in defining the specificity of PDZ domains, a set of energetically coupled positions for the binding site was identified, including unexpected long-range interactions. Whilst extremely interesting, however, this study was subsequently reassessed and it was found that the energetic coupling was not a special feature of the coevolved network of residues in PDZ domains [48,49]. In addition, as mentioned, in the context of the PDZ3-SH3-GK supramodule the experimentally determined allosteric network for PDZ3 is different from that in isolation [50]. Thus, statistical coupling from sequence analysis is not necessarily a reporter of energetic coupling and allostery and must, therefore, be supported with extensive experimental data.

7. Double Mutant Cycles by Native Mass Spectrometry

A very interesting application of double mutant cycles has been recently implemented by studying the formation of protein-protein interactions via native mass spectrometry [51]. The intrinsic power of this technique lies in the ability to transfer protein complexes to the gas phase, without altering their native state [52–54]. Hence, by measuring a single high-resolution native mass spectrum and determining the intensities of the complexes formed by the two wild-type proteins, the complex of each wild-type protein with a mutant protein, and the complex of the two mutant proteins, it is possible to obtain the pairwise interaction energies between the two mutated residues with great precision. In fact, it can be demonstrated that native mass spectrometry both circumvents the determination of individual binding constants, which are of course prone to experimental error, and is independent of the concentration of the free unbound species [55]. Remarkably, the method may

be employed directly on crude cell lysates [56], which further simplifies the experiments and interpretations. In fact, in these cases, the purification of the mutant variants may not be required, which simplifies the experimental approach consistently.

Double mutant cycles by mass spectrometry have been successfully employed to study the binding between the proteins E9 and Im2 [51] and, more recently, to understand the dimerization of SOD1 in naturally-occurring mutants associated with ALS. The authors analyzed the coupling constants by co-expressing wild-type SOD1 with ALS-causing mutations [57]. The result is a quantitative measurement of the inherent preference of these variants to form homo-dimers (mutant-mutant or wild type-wild type) rather than heterodimers (mutant-wild type). Of exceptional interest, the analysis successfully demonstrated that heterodimerization preference of SOD1 in ALS-causing mutations correlates with the reported average duration of the disease. Their findings suggest that heterodimerization of mutant variants of SOD1 is directly involved in the development of ALS and provide another clear example of how the initial events of self-assembly represent a critical point in the intimate link between protein misfolding and disease [58].

8. Conclusions

The double mutant cycles methodology represents a very powerful technique offering the tantalizing possibility to address directly the strength of pairwise interactions in proteins. As exemplified in the case of PDZ domains, the analysis of the sign of the observed energetic coupling may be used to describe protein selectivity in conserved homologous protein-protein interaction domains. Furthermore, in conjunction with other techniques, it may represent a future tool to address ambitious tasks, such as the understanding of structure of metastable states as well as the description of aggregation phenomena. With the specific aim to help the experimentalist, we have recapitulated here the key features, along with the associated caveats and pitfalls, as well as some key examples of applications.

Author Contributions: L.P., P.J., and S.G. wrote the first version of the manuscript. L.P., P.J., A.T., F.M., L.V. and S.G. discussed and revised the manuscript. L.P., P.J., A.T., F.M., L.V. and S.G. All authors have read and agreed to the published version of the manuscript.

Funding: Work partly supported by grants from Sapienza University of Rome (B52F16003410005, RP11715C34AEAC9B and RM1181641C2C24B9 to S.G), the Associazione Italiana per la Ricerca sul Cancro (individual grant, IG 2020, 24551, to S.G.), the Istituto Pasteur Italia (Teresa Ariaudo Research Project 2018, to A.T.) and the Swedish Research Council (2020-04395 to P.J.). F.M. is the recipient of a FIRC-AIRC fellowship.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript or in the decision to publish the results.

References

1. Yip, K.M.; Fischer, N.; Paknia, E.; Chari, A.; Stark, H. Atomic-resolution protein structure determination by cryo-EM. *Nature* **2020**, *587*, 157–161. [CrossRef] [PubMed]
2. Wüthrich, K. Protein structure determination in solution by nmr spectroscopy. *World Sci. Res.* **1995**, *5*, 11–14.
3. Dobson, C.M. Biophysical techniques in structural biology. *Annu. Rev. Biochem.* **2019**, *88*, 25–33. [CrossRef] [PubMed]
4. Fersht, A.R. *Structure and Mechanism in Protein Science*; Freeman: New York, NY, USA, 1999.
5. Horovitz, A.; Fleisher, R.C.; Mondal, T. Double-mutant cycles: New directions and applications. *Curr. Opin. Struct. Biol.* **2019**, *58*, 10–17. [CrossRef]
6. Horovitz, A. Double-mutant cycles: A powerful tool for analyzing protein structure and function. *Fold. Des.* **1996**, *1*, 121–126. [CrossRef]
7. Cockroft, S.L.; Hunter, C.A. Chemical double-mutant cycles: Dissecting non-covalent interactions. *Chem. Soc. Rev.* **2007**, *36*, 172–188. [CrossRef]
8. Sali, D.; Bycroft, M.; Fersht, A.R. Surface electrostatic interactions contribute little of stability of barnase. *J. Mol. Biol.* **1991**, *220*, 779–788.
9. Otzen, D.E.; Fersht, A.R. Analysis of protein-protein interactions by mutagenesis: Direct versus indirect effects. *Protein. Eng.* **1999**, *12*, 41–45. [CrossRef]
10. Bateson, W. *Mendel's Principles of Heredity: A Defence*; Cambridge University Press: Cambridge, UK, 1902.

11. Moore, J.H. A global view of epistasis. *Nat. Genet.* **2005**, *37*, 13–14. [CrossRef]
12. Starr, T.N.; Thornton, J.W. Epistasis in protein evolution. *Protein Sci.* **2016**, *25*, 1204–1218. [CrossRef]
13. Winter, G.; Fersht, A.R.; Wilkinson, A.J.; Zoller, M.; Smith, M. Redesigning enzyme structure by site-directed mutagenesis: Tyrosyl tRNA synthetase and ATP binding. *Nature* **1982**, *299*, 756–758. [CrossRef] [PubMed]
14. Carter, P.J.; Winter, G.; Wilkinson, A.J.; Fersht, A.R. The use of double mutants to detect structural changes in the active site of the tyrosyl-tRNA synthetase (Bacillus stearothermophilus). *Cell* **1984**, *38*, 835–840. [CrossRef]
15. Fersht, A.R.; Sato, S. Φ-Value analysis and the nature of protein-folding transition states. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 7976–7981. [CrossRef] [PubMed]
16. Horovitz, A.; Fersht, A.R. Strategy for analysing the co-operativity of intramolecular interactions in peptides and proteins. *J. Mol. Biol.* **1990**, *214*, 613–617. [CrossRef]
17. Fersht, A.R.; Matouschek, A.; Serrano, L. The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.* **1992**, *224*, 771–782. [CrossRef]
18. Serrano, L.; Horovitz, A.; Avron, B.; Bycroft, M.; Fersht, A.R. Estimating the contribution of engineered surface electrostatic interactions to protein stability by using double-mutant cycles. *Biochemistry* **1990**, *29*, 9343–9352. [CrossRef]
19. Sato, S.; Religa, T.L.; Fersht, A.R. Φ-Analysis of the folding of the b domain of protein a using multiple optical probes. *J. Mol. Biol.* **2006**, *360*, 850–864. [CrossRef] [PubMed]
20. Serrano, L.; Bycroft, M.; Fersht, A.R. Aromatic-aromatic interactions and protein stability: Investigation by double-mutant cycles. *J. Mol. Biol.* **1991**, *218*, 465–475. [CrossRef]
21. Horovitz, A.; Serrano, L.; Avron, B.; Bycroft, M.; Fersht, A.R. Strength and co-operativity of contributions of surface salt bridges to protein stability. *J. Mol. Biol.* **1990**, *216*, 1031–1044. [CrossRef]
22. Vaughan, C.K.; Harryson, P.; Buckle, A.M.; Fersht, A.R. A structural double-mutant cycle: Estimating the strength of a buried salt bridge in barnase. *Acta Crystallogr. D* **2002**, *58*, 591–600. [CrossRef]
23. Green, S.M.; Shortle, D. Patterns of nonadditivity between pairs of stability mutations in staphylococcal nuclease. *Biochemistry* **1993**, *32*, 10131–10139. [CrossRef] [PubMed]
24. Chen, J.; Stites, W.E. Energetics of side chain packing in staphylococcal nuclease assessed by systematic double mutant cycles. *Biochemistry* **2001**, *40*, 14004–14011. [CrossRef] [PubMed]
25. Marqusee, S.; Sauer, R.T. Contributions of a hydrogen bond/salt bridge network to the stability of secondary and tertiary structure in lambda repressor. *Protein Sci.* **1994**, *3*, 2217–2225. [CrossRef] [PubMed]
26. Myers, J.K.; Oas, T.G. Contribution of a buried hydrogen bond to λ repressor folding kinetics. *Biochemistry* **1999**, *38*, 6761–6768. [CrossRef] [PubMed]
27. Thakur, J.K.; Yadav, A.; Yadav, G. Molecular recognition by the KIX domain and its role in gene regulation. *Nucleic Acids Res.* **2014**, *42*, 2112–2125. [CrossRef] [PubMed]
28. Troilo, F.; Bonetti, D.; Toto, A.; Visconti, L.; Brunori, M.; Longhi, S.; Gianni, S. The folding pathway of the KIX domain. *J. Am. Chem. Soc.* **2017**, *12*, 1683–1690. [CrossRef]
29. Myers, J.K.; Pace, C.N.; Scholtz, J.M. Denaturant m values and heat capacity changes: Relation to changes in accessible surface areas of protein unfolding. *Protein Sci.* **1995**, *4*, 2138–2148. [CrossRef]
30. Scaloni, F.; Gianni, S.; Federici, L.; Falini, B.; Brunori, M. Folding mechanism of the C-terminal domain of nucleophosmin: Residual structure in the denatured state and its pathophysiological significance. *Fed. Proc.* **2009**, *23*, 2360–2365. [CrossRef]
31. Perutz, M.F.; Kendrew, J.C.; Watson, H.C. Structure and function of haemoglobin: II. Some relations between polypeptide chain configuration and amino acid sequence. *J. Mol. Biol.* **1965**, *13*, 669–678. [CrossRef]
32. Goodey, N.M.; Benkovic, S.J. Allosteric regulation and catalysis emerge via a common route. *Nat. Chem. Biol.* **2008**, *4*, 474–482. [CrossRef]
33. Schreiber, G.; Fersht, A.R. Rapid, electrostatically assisted association of proteins. *Nat. Struct. Biol.* **1996**, *3*, 427–431. [CrossRef] [PubMed]
34. Ming, D.; Chen, R.; Huang, H. Amino-acid network clique analysis of protein mutation non-additive effects: A case study of lysozme. *Int. J. Mol. Sci.* **2018**, *19*, 1427. [CrossRef] [PubMed]
35. Rajasekaran, N.; Sekhar, A.; Naganathan, A.N. A universal pattern in the percolation and dissipation of protein structural perturbations. *J. Phys. Chem. Lett.* **2017**, *8*, 4779–4784. [CrossRef] [PubMed]
36. Horovitz, A. Non-additivity in protein-protein interactions. *J. Mol. Biol.* **1987**, *196*, 733–735. [CrossRef]
37. Nussinov, R.; Tsai, C.-J. Allostery without a conformational change? Revisiting the paradigm. *Curr. Opin. Struct. Biol.* **2015**, *30*, 17–24. [CrossRef]
38. Schreiber, G.; Fersht, A.R. Energetics of protein-protein interactions: Analysis of the Barnase-barstar interface by single mutations and double mutant cycles. *J. Mol. Biol.* **1995**, *248*, 478–486. [CrossRef]
39. Ibarra-Molero, B.; Zitzewitz, J.A.; Matthews, C.R. Salt-bridges can stabilize but do not accelerate the folding of the homodimeric coiled-coil peptide GCN4-p1. *J. Mol. Biol.* **2004**, *336*, 989–996. [CrossRef]
40. Jemth, P.; Gianni, S. PDZ domains: Folding and binding. *Biochemistry* **2007**, *46*, 8701–8708. [CrossRef]
41. Gianni, S.; Haq, S.R.; Montemiglio, L.C.; Jürgens, M.C.; Engström, Å.; Chi, C.N.; Brunori, M.; Jemth, P. Sequence-specific long range networks in PSD-95/Discs Large/ZO-1 (PDZ) domains tune their binding selectivity. *J. Biol. Chem.* **2011**, *286*, 27167–27175. [CrossRef]

42. Hultqvist, G.; Haq, S.R.; Punekar, A.S.; Chi, C.N.; Engström, Å.; Bach, A.; Strømgaard, K.; Selmer, M.; Gianni, S.; Jemth, P. Energetic pathway sampling in a protein interaction domain. *Structure* **2013**, *21*, 1193–1202. [CrossRef]
43. Eildal, J.N.N.; Hultqvist, G.; Balle, T.; Stuhr-Hansen, N.; Padrah, S.; Gianni, S.; Strømgaard, K.; Jemth, P. Probing the role of backbone hydrogen bonds in protein-peptide interactions by amide-to-ester mutations. *J. Am. Chem. Soc.* **2013**, *135*, 12998–13007. [CrossRef] [PubMed]
44. Malagrinò, F.; Troilo, F.; Bonetti, D.; Toto, A.; Gianni, S. Mapping the allosteric network within a SH3 domain. *Sci. Rep.* **2019**, *9*, 8279. [CrossRef] [PubMed]
45. Laursen, L.; Kliche, J.; Gianni, S.; Jemth, P. Supertertiary protein structure affects an allosteric network. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 24294–24304. [CrossRef] [PubMed]
46. Horovitz, A.; Bochkareva, E.S.; Yifrach, O.; Girshovich, A.S. Prediction of an inter-residue interaction in the chaperonin groel from multiple sequence alignment is confirmed by double-mutant cycle analysis. *J. Mol. Biol.* **1994**, *238*, 133–138. [CrossRef]
47. Lockless, S.W.; Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **1999**, *286*, 295–299. [CrossRef] [PubMed]
48. Chi, C.N.; Elfstrom, L.; Shi, Y.; Snall, T.; Engstrom, A.; Jemth, P. Reassessing a sparse energetic network within a single protein domain. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 4679–4684. [CrossRef]
49. Fodor, A.A.; Aldrich, R.W. On evolutionary conservation of thermodynamic coupling in proteins. *J. Biol. Chem.* **2004**, *279*, 19046–19050. [CrossRef] [PubMed]
50. Gautier, C.; Troilo, F.; Cordier, F.; Malagrinò, F.; Toto, A.; Visconti, L.; Zhu, Y.; Brunori, M.; Wolff, N.; Gianni, S. Hidden kinetic traps in multidomain folding highlight the presence of a misfolded but functionally competent intermediate. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 19963–19969. [CrossRef]
51. Sokolovski, M.; Cveticanin, J.; Hayoun, D.; Korobko, I.; Sharon, M.; Horovitz, A. Measuring inter-protein pairwise interaction energies from a single native mass spectrum by double-mutant cycle analysis. *Nat. Commun.* **2017**, *8*, 212. [CrossRef]
52. Chandler, S.A.; Benesch, J.L. Mass spectrometry beyond the native state. *Curr. Opin. Chem. Biol.* **2018**, *42*, 130–137. [CrossRef]
53. Mehmood, S.; Allison, T.M.; Robinson, C.V. Mass spectrometry of protein complexes: From origins to applications. *Annu. Rev. Phys. Chem.* **2015**, *66*, 453–474. [CrossRef] [PubMed]
54. Lössl, P.; van de Waterbeemd, M.; Heck, A.J. The diverse and expanding role of mass spectrometry in structural and molecular biology. *EMBO Rep.* **2016**, *35*, 2634–2657. [CrossRef] [PubMed]
55. Sharon, M.; Horovitz, A. Probing allosteric mechanisms using native mass spectrometry. *Curr. Opin. Struct. Biol.* **2015**, *34*, 7–16. [CrossRef]
56. Cveticanin, J.; Netzer, R.; Arkind, G.; Fleishman, S.J.; Horovitz, A.; Sharon, M. Estimating interprotein pairwise interaction energies in cell lysates from a single native mass spectrum. *Anal. Chem.* **2018**, *90*, 10090–10094. [CrossRef] [PubMed]
57. Cveticanin, J.; Mondal, T.; Meiering, E.M.; Sharon, M.; Horovitz, A. Insight into the autosomal-dominant inheritance pattern of SOD1-associated ALS from native mass spectrometry. *J. Mol. Biol.* **2020**, *432*, 5995–6002. [CrossRef] [PubMed]
58. Chiti, F.; Dobson, C.M. Protein misfolding, amyloid formation, and human disease: A summary of progress over the last decade. *Annu. Rev. Biochem.* **2017**, *86*, 27–68. [CrossRef]