Università degli Studi di Milano-Bicocca

Dipartimento di Biotecnologie e Bioscienze

Dottorato di ricerca in Biotecnologie e Biologia

XXX Ciclo

Effects of electrostatic charges on aggregation and conformation of intrinsically disordered proteins

Giulia Tedeschi

Anno Accademico 2016/2017

SCUOLA DI DOTTORATO

UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Dipartimento di / Department of Department of Biotechnology and Biosciences

Dottorato di Ricerca in / PhD program in Biotechnology and Biology

Ciclo / Cycle XXX

Curriculum in Biotechnology

Effects of electrostatic charges on aggregation and conformation of intrinsically disordered proteins

Cognome / Surname: Tedeschi Nome / Name Giulia

Matricola / Registration number: 730533

Tutore / Tutor: Prof. Stefania Brocca

Coordinatore / Coordinator: Prof. Marco Ercole Vanoni

Anno Accademico 2016/2017

Preface

Three years ago, when I joined Stefania Brocca's team at the Department of Biotechnology and Biosciences (BTBS) of Uni-MiB, my "great love" was molecular biology and genetics. I had no much feeling for biochemistry, and poor knowledge on protein structure, and even less on intrinsically disordered proteins, namely IDPs.

Today I can say that my PhD experience made me to change. And not only from a scientific point of view. I was motivated to "learn fast" and besides reading books and articles, I had the opportunity to attend seminars, conferences and Summer Schools around Europe. I've travelled also to reach the CNRS of Marseille where, from January to July 2017, I visited Sonia Longhi's lab at AFMB (Architecture et Fonction des Macromolécules Biologiques). There, I've learnt new techniques and completed my experimental work.

Finally, from August to December 2017, I was engaged in writing my dissertation to which I've dedicated a lot of energy. Indeed, I've intended to make it understandable also to "laypersons", like me in 2014.

My dissertation begins with an "Introduction" where I describe general aspects of IDPs and their propensity to aggregate and to collapse. The "Introduction" is completed by an "Appendix", which deals with genes encoding IDPs. This topic represents a new knowledge frontier in the field of IDPs and its inclusion in this thesis reflects my genetics background. The second chapter, "Methods", does not contain any protocol, instead it presents in simple terms theoretical aspects of bioinformatics, biochemical and biophysical methods used during my work. The third chapter explains the aims of the project, together with main results and brief, general conclusions. Through this chapter, readers may have an overall, not-fragmented picture of the entire work. The fourth chapter, "Experimental work" encloses the two manuscripts in which converged the main results of my project.

More in detail, the work entitled "Aggregation properties of a disordered protein are tunable by pH and depend on its net charge per residue" has been carried out

during the first and second year of my PhD course, being entirely performed at the BTBS Department, with the tutoring of Stefania Brocca and Marina Lotti. The work was made possible thanks to the collaboration with Antonino Natalello that contributed with infrared spectroscopy analyses. My colleagues Marco Mangiagalli and Sara Chmielewska helped with the production and biochemical analyses of proteins. This work has already been published in October 2017, in BBA General Subject (https://doi.org/10.1016/j.bbagen.2017.09.002).

The second work, entitled "Clustering of charged residues and proline content affect conformational properties of intrinsically disordered proteins", has been carried out during the last year of my PhD course. The work has been planned and partially performed at the BTBS Department, under the supervision of Stefania Brocca and thanks to the collaboration with Rita Grandori and Carlo Santambrogio, experts in electrospray ionization mass spectrometry. A large part of this work has been carried out at the CNRS-AFMB, with the tutoring of Sonia Longhi. A relevant part of biophysical analyses was carried out at the Grenoble Synchrotron with the help of Edoardo Salladini, PhD student at the AFMB of Marseille. This second work too is going to be submitted for publication.

In conclusion, I'm convinced that the goals represented by my scientific results and the preparation of this thesis would have never been reached without my tutors Stefania Brocca, Marina Lotti, Sonia Longhi and all my colleagues that constantly gave me the possibility to grow up from a personal and scientific point of view.

I hope you can appreciate my work and enjoy your reading.

Milano, February 10, 2018

Giulia Tedeschi

Index

Abbreviations

Abbreviations

AS: alternative splicing

CD: circular dichroism

CH plot: charge- hydropathy plot

CIDER: classification of Intrinsically Disordered Ensemble Regions

DLS: dynamic light scattering

D_{max}: maximal intramolecular distance

EOM: ensemble optimization method

FCR: fraction of charged residues

FT-IR: Fourier transform infrared spectroscopy

GFP: green fluorescent protein

IDP: intrinsically disordered protein

IDR: intrinsically disordered region

LM: linear motif

MG: molten globule

MM: molecular mass

MW: molecular weight

MORF: molecular recognition features

NCPR: net charge per residues

NF: native folded

N_{TAIL}: C-terminal domain of measles virus nucleoprotein

PDB: protein Data Bank

pI: isoelectric point

PMG: pre-molten globule

PNT: N-*terminus* moiety of measles virus phosphoprotein

PNT4: residues 300-404 of Hendra virus nucleoprotein

PONDR-FIT: predictor of natural disordered region

P(r) plot: pair distribution plot

PSE: preformed structural element

PTM: post-translational modification

RC: random coil

Rg: gyration radius

RNP: ribonucleoprotein

Rs: Stokes radius

SAXS: small angle X-ray scattering

SDS: sodium dodecyl sulphate

SEC: size exclusion chromatography

Abstract

"Intrinsic disorder" is generally referred to the conformational status of native proteins lacking secondary and/or tertiary structure, although not exposed to any denaturing agent. These proteins, which are called intrinsically disordered (IDP/IDRs), represent a large class in the proteomes of all living beings, with a remarkable abundance among viruses and more complex eukaryotes.

IDPs have been recognized to be involved in many relevant physiological and pathological functions, such as the condensation into membrane-less organelles or the fibrillation in amyloid bodies. It is becoming clearer that fast and massive intermolecular interactions involving IDPs are governing both kinds of phenomena and that pathologies can arise from dysregulations of conformational properties and aggregation ability.

The conformation and aggregation features of IDPs have been ascribed in turn to several factors, such as sequence length, hydrophobic interactions, hydrogen bonds or electrostatic charges. The latter deserves particular attention since charged residues are particularly abundant in IDPs. The net charge per residue (NCPR), the total fraction of charged residues (FCR), and the linear distribution of opposite charges (κ value) have been recently regarded as the primary determinants of IDPs conformational properties.

The first part of the experimental work presented in this thesis was inspired by the concept of NCPR, which represents the net charge normalized by the protein length. The aim is to describe how the NCPR influences the ability of IDPs to respond to environment pH changes through loss of solubility. The N-terminal domain of phosphoprotein (PNT) from measles virus was used as a model IDP. Moreover, the wild type (*wt*) protein was compared with some PNT variants designed to share same hydrophobicity and FCR, but differing in NCPR and isoelectric points (pI). Tested proteins showed a solubility minimum close to their pI, as expected, and a pH-dependent decrease of solubility not equal, but driven by the NCPR of each variant. Our data suggest that the overall solubility of a

protein can be dictated by some protein regions prompter to respond to pH changes.

The second part of experimental work was inspired by the concept of charge clustering. It was aimed at verifying that the compaction properties of IDPs are tunable by the κ value. We have used two well-characterized IDPs, namely measles virus nucleoprotein C-terminal region (N_{TAIL}) and Hendra virus PNT4, as model systems. Taking advantage of the high sequence *designability* of IDPs, genes of PNT4 and N_{TAIL} were redesigned to obtain two sets of synthetic proteins each including the *wt* form and two "κ variants". In low-κ variants, charged amino acids are most evenly distributed, in high-κ variants charges are clustered as much as possible at the N- and C-termini. All κ variants, along with *wt* forms, were subjected to various biophysical and biochemical techniques to assess their conformational properties. Overall, experimental data confirm the expected trend, with compactness increasing with κ value. The increase of compactness does not follow a general trend, but it is protein-specific and related to the proline content. All together, these findings confirm previous theoretical and experimental data on the role of charged residues frequency (NCPR) and distribution (κ). The main value of this experimental work is in pinpointing the context, which is the environment – pH – or the amino acid composition – proline % –, where such driving forces of aggregation and compaction are mostly effective. This knowledge is useful not only to describe how the conformational behavior of IDPs is encoded by their amino acid sequence, but also to rationally design non-natural IDPs with desired conformational and aggregation properties.

Riassunto

"Intrinsecamente disordinata" viene definita una proteina nativa priva di struttura secondaria o terziaria, non esposta ad agenti denaturanti. Le proteine con queste caratteristiche sono indicate come IDP/IDR, acronimo dall'inglese "*intrinsically disordered protein/region*" e rappresentano una ampia porzione del proteoma di tutti gli esseri viventi ed in particolare di virus ed eucarioti superiori.

Le IDP sono coinvolte in molte funzioni fisiologiche e patologiche, come la condensazione in organuli cellulari privi di membrane e la formazione di fibrille associate ad amiloidosi. Entrambi questi fenomeni sono associati alla capacità delle IDP di formare interazioni intermolecolari. Stati patologici possono essere causati da disfunzioni e cattiva regolazione delle proprietà conformazionali e di aggregazione delle IDP.

L'aggregazione e la conformazione delle IDP sono state ascritte a diversi fattori: la lunghezza della catena amminoacidica, le interazioni idrofobiche, i legami ad idrogeno e le cariche elettrostatiche. A questa ultima abbiamo rivolto la nostra attenzione dal momento che le IDP sono ricche di amminoacidi carichi. Più recentemente, la carica netta per residuo (NCPR) e la frazione totale di residui carichi (FCR), così come la distribuzione di residui di carica opposta (valore κ) sono stati considerati i principali determinanti della conformazione delle IDP.

La prima parte del lavoro sperimentale presentato riguarda il concetto di NCPR, cioè la carica netta normalizzata per la lunghezza della proteina. L'obiettivo è di descrivere come questo parametro influenzi la capacità delle IDP di rispondere a cambiamenti di pH con conseguente perdita di solubilità. Come modello è stata utilizzata la regione N-terminale della proteina P (PNT) del virus del morbillo ed a partire da questa è stata ottenuta una serie di varianti dotate della stessa idrofobicità ed FCR, ma differente NCPR e punto isoelettrico (pI). Le proteine analizzate mostrano solubilità minima in corrispondenza del loro valore di pI, come atteso. La perdita di solubilità dipendente da pH non avviene per tutte in ugual misura, ma è guidata dal valore di NCPR di ciascuna variante proteica. I

dati sperimentali suggeriscono come la solubilità complessiva di una proteina possa essere legata al suo valore di NCPR e da questo dipenda la risposta a variazioni di pH.

La seconda parte del lavoro sperimentale si è ispirata al concetto di *clusterizzazione* di cariche ed ha come obiettivo la valutazione di come le proprietà di compattezza delle IDP dipendano dal valore di κ. In questo caso sono state utilizzate due IDP ben caratterizzate, la regione C-terminale della proteina N (N_{TAIL}) dal virus del morbillo e PNT4 da Hendra virus. Grazie alla possibilità di modificare la sequenza amminoacidica delle IDP senza interferire sul complessivo disordine strutturale, entrambi i geni sono stati riprogettati. Sono stati ottenuti due set di proteine sintetiche, ciascuno contenente una proteina *wild type* (*wt*) e due varianti in cui le cariche sono uniformemente distribuite (*low* κ) o completamente segregate all'N- ed al C-*terminus* (*high* κ). Le proprietà conformazionali della proteina *wt* e delle corrispondenti varianti sono state valutate mediante tecniche biofisiche e biochimiche. Complessivamente i dati sperimentali confermano l'andamento atteso, cioè un aumento del grado di compattezza conformazionale all'aumentare dei valori di κ, secondo una proporzione che è tipica di ciascuna proteina in relazione al suo contenuto di proline.

Complessivamente i risultati ottenuti confermano precedenti dati computazionali e sperimentali, suggerendo come residui carichi, attraverso la loro frequenza (NCPR) e distribuzione (κ), influenzino solubilità e compattezza delle IDP. I due lavori sperimentali sottolineano l'importanza del contesto, ambientale (ad esempio, le condizioni di pH) o di sequenza (la percentuale di proline), sull'efficacia di NCPR e della distribuzione di carica come determinanti di solubilità e compattezza conformazionale delle IDP. La rilevanza di queste informazioni è legata non solo allo studio IDP naturali, ma anche alla progettazione razionale di proteine non naturali con proprietà aggregative e conformazionali ben definite.

1.Introduction

1.1 General features of IDPs

The aim of this chapter is to introduce IDPs and to highlight their uniqueness under compositional and structural aspects.

The conformational and compositional peculiarities of IDPs will be described, along with their propensity to be the target of post-translational modifications (PTMs), to exhibit promiscuous function, to participate in interaction hubs and phase transition phenomena.

Recent studies indicate that the peculiarities of IDPs are reflected also at the level of their genes, splicing mechanisms and translation. These topics are presented in the Appendix 1.

1.1.1 From structural to "unstructural" biology

Starting from 1970, more than ten thousand of structures have been solved and deposited in Protein Data Bank (PDB). The occurrence of these data supports the "structure-function paradigm" stating that a protein function stems from its well-defined structure. More recently, starting from the '90, the scenario has changed because of the discovery of a new class of proteins devoid of a defined three-dimensional (3-D) structure and yet able to exert their biological functions. The present name of "intrinsically disordered proteins" (IDPs) or "intrinsically disordered regions" (IDRs) used to indicate them just refers to the lack of a well-defined secondary and/or tertiary structure, and to the fact that this property occurs under physiological conditions (Dunker et al., 2013). The existence of IDPs or IDRs does not only "break the rule" of structural biology dogma (Dunker et al., 2001a), but has also allowed to answer some questions remained open for several years: "what does account for the missing electron density in PDB structures?", "why are some proteins so sensitive to proteolysis?", "why do some proteins possess a particular behaviour in size exclusion chromatography, or gel electrophoresis?" (Habchi et al., 2014). Nowadays, "unstructural" biology

involves many scientists fascinated by the chance to unveil IDP secrets. As a result, the number of structurally and functionally characterised IDPs is growing rapidly, together with the number of papers on IDPs (**Figure 1.1**) (Uversky, 2014).

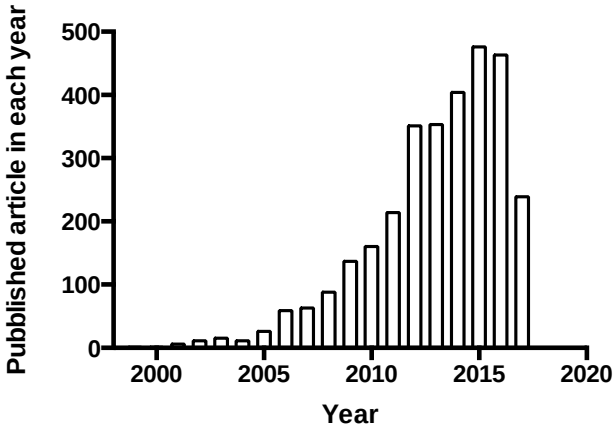


Figure 1.1. The number of publications related to IDPs by year, from 1999 to 2017 witnesses the growing interest of researchers in "unstructural" biology (PUBMED, July 2017).

Since its conceptual raise, the new class of IDPs and IDRs has brought the need to reconsider the pre-existing schemes of structural categorization. One of the first theory including the concept of structural disorder is the so-called "protein quartet model". It proposes that protein function can arise from four types of conformational states and thereof transitions: random coil (RC), pre-molten globule (PMG), molten globule (MG) and folded state (**Figure 1.2**).

Unbound, disordered regions could fall into all categories except the "folded state". The PMG state represents a "squeezed" and partially ordered form of the coil with some residual secondary structure. The MG state is a collapsed disordered form in which native secondary structure exists although the protein lacks a well-packed core. Finally, the RC shows little or no secondary structure.

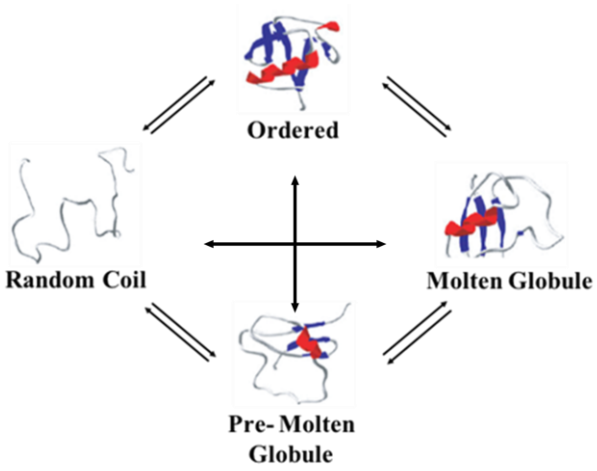


Figure 1.2. The protein quartet model of protein conformational states. Protein function arises from four types of conformation of polypeptide chain and transition between any of these states (adapted from Van Der Lee et al., 2014 and Habchi et al., 2014).

More recently, the concept of "conformational continuum" has been proposed to include the wide repertoire of documented protein conformations, ranging from fully structured to completely disordered states (**Figure 1.3**) (Uversky and Dunker, 2010).

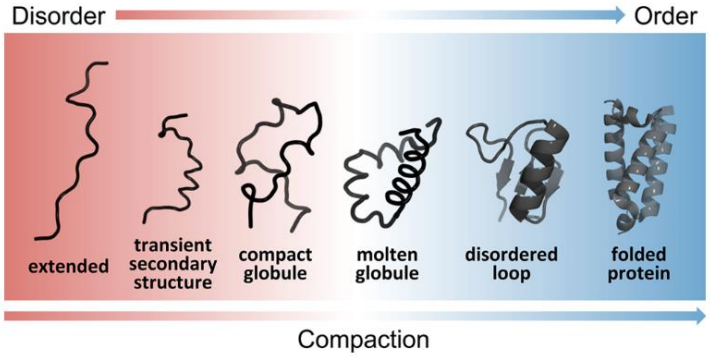


Figure 1.3. Schematic representation of structural disorder continuum, ranging from highly dynamic, expanded conformational ensembles (left, red) to compact, dynamically restricted, fully folded proteins (right, blue) (Van Der Lee et al., 2014).

1.1.2 Amino acid composition of IDPs

In comparison with structured ("globular") proteins, IDPs show a peculiar amino acid composition (**Figure 1.4**) and further, they are characterised by repeats of low-complexity sequence (Uversky, 2011).

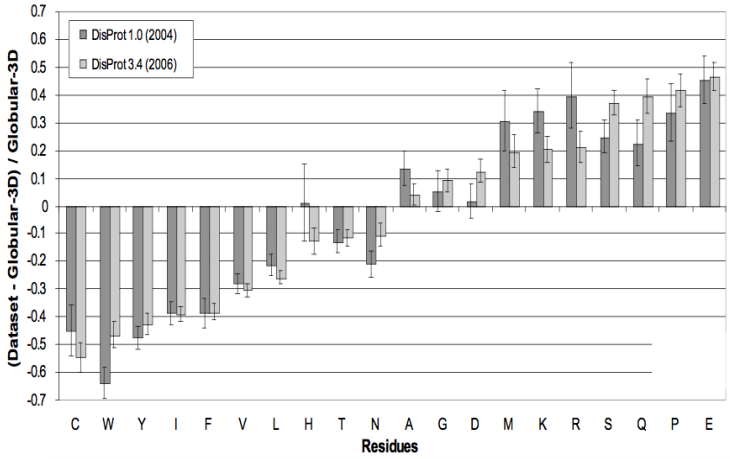


Figure 1.4. Amino acid composition of two sets of IDPs (Disprot 1.0 and Disprot 3.4), relative to a set of globular proteins (PDB 3D) (Dunker et al., 2008)

IDPs are depleted in so-called order-promoting amino acids such as Ile, Leu, Val, Trp, Tyr, Phe, Cys and Asn, and rich of disorder-promoting amino acids, Ala, Arg, Gly, Gln, Ser, Glu, Lys and Pro (Dunker et al., 2008) (**Table 1.1**).

Amino acid	Globular protein composition (%)	Disordered protein composition (%)
Gly(G)	7.4	4.3
Ala (A)	7.9	7.2
Leu (L)	8.9	5.4
Ile (I)	5.6	3.7
Met (M)	2.2	1.3
Phe (F)	4.0	1.7
Trp (W)	1.4	0.3
Val (V)	6.8	8.0
Pro (P) *	4.7	12.0
Cys (C)	1.6	0.6
Ser (S)	6.7	6.9
Thr (T)	5.9	5.1
Asn (N)	4.5	2.0
Gln (Q)	3.8	4.5
Tyr (Y)	3.4	1.4
Arg (R)	4.9	4.2
His (H)	2.3	1.5
Lys (K)*	6.3	10.4
Asp (D)	5.5	5.0
Glu (E)*	6.2	14.3

Table 1.1 Amino acid frequencies (%) in globular proteins (SwissProt) and IDPs (DisEMBL). Asterisks indicate most abundant amino acids in IDPs (Dunker et al., 2002).

Disordered-promoting residues represent the 64% in IDPs, while they are only the 48% in globular proteins. Pro, Lys and Glu are at least two times more frequent in IDPs than in globular proteins (marked with asterisks in **Table 1.1**). Overall, with respect to globular proteins, the absolute content of hydrophobic residues does not change, while the percentage of charged ones is sometimes drastically increased. Since hydrophobic residues mainly contribute to hydrophobic core, one can hypothesize most IDPs can retain some "seed of order" or compactness, coupled with great flexibility. Overall, the combination of high charge and low hydrophobicity has been considered to cause high flexibility (Habchi and Longhi, 2012), and solubility (Uversky and Longhi, 2011) of IDPs. These features can account for the general sensitivity of IDPs to environment conditions (see chapter 2.4).

1.1.3 IDPs as polyampholytes

High frequency of positively and negatively charged groups makes the definition of "polyampholytes" suitable for IDPs. Besides net charge and isoelectric point (pI), their properties have been described through various parameters such as net charge per residue (NCPR) (Mao et al., 2010), total fraction of charged residues (FCR), and linear distribution of opposite charges (κ value) (Das and Pappu, 2013).

The NCPR determines the dimension of unfolded polypeptides, as theorized for polyelectrolyte (Mao et al., 2010; Marsh and Forman-Kay, 2010; Müller-Späth et al., 2010). The higher NCPR, the more extended the protein size, which can be described by hydrodynamic radius, gyration radius, the mean end-to-end distance etc. This behaviour depends from the favourable free energies of solvation of charged sidechains and from the electrostatic repulsions among like-charge residues. Hence, high NCPR-proteins are likely to have a random coil conformation. More in general, the type of conformation of a given IDP might be

predicted on its NCPR value. Based on this reasoning, Mao et al. (2010) have annotated a subset of IDP sequences from the DisProt database (Sickmeier et al., 2006) by using a predictive diagram of states (Mao et al., 2010). However, NCPR alone is not a sufficient descriptor of conformational properties, due to the fact they are polyelectrolytes. Hence, by using the lone net charge, or NCPR, one risks overlooking the importance of FCR value and the effects of linear distribution of charges. Patterning of positively and negatively charged residues can influence not only the global dimensions, but also the amplitudes of conformational fluctuations (Das and Pappu, 2013; Holehouse and Pappu, 2018). A way to take into account of charge patterning consists in the calculation of the parameter κ, which is related to the linear distribution of charges along a protein sequence (Das and Pappu, 2013). The κ value ranges from 0 to 1. In the case of evenly distributed positive and negative charges, κ value is null. On the contrary, for charges segregated in two distinct clusters, κ reaches its maximum value (*i.e.*, 1). *In-silico* studies on simple polypeptides composed of Glu and Lys indicate that when $\kappa \rightarrow 0$, electrostatic repulsions and attractions within the chain are counterbalanced, leading to a self-avoiding random walk or generic Flory-type random coil conformational state. When oppositely charged residues are segregated within the sequence ($\kappa \rightarrow 1$), hairpin-like conformations emerge because long-range electrostatic attractions are preferred (Das and Pappu, 2013). Thus, an inverse correlation exists between κ value and gyration *radii* of IDPs and unfolded proteins (Das et al., 2015).

The direct dependence of protein compactness from the value of κ has been proved by experimental data obtained on permutant synthetic IDRs derived from natural proteins such as the Notch receptor (Sherry et al., 2017) and p27 (Das et al., 2016).

1.1.4 Post-translational modifications of IDPs

Post-translational modifications (PTMs) mainly consist of enzymatic addition/modification of chemical groups in the primary structure of a protein. They can occur at any stage of protein's lifetime, often providing key regulatory mechanisms in different biological processes.

Due to their solvent accessibility, IDPs are easily accessible to modifying enzymes. Good evidence of this can be found considering that PTM-catalysing enzymes in eukaryotic cells preferentially recognize IDRs as target (Uversky, 2013b; Uversky and Dunker, 2010). Moreover, IDPs are suitable to receive multiple functional groups in a relatively narrow sequence segment. Indeed, conformational adjustments can effectively compensate for steric hindrance of bulky groups and repulsive forces among charged moieties.

Overall, PTMs may significantly expand the functional versatility of IDPs through a range of structural changes, including disorder-to-order transitions (Babu et al., 2012; Xie et al., 2007).

The most common PTMs is phosphorylation: at least 75% of eukaryotic proteins may be phosphorylated, and most phosphorylation sites are within IDPs or IDRs. In particular, phosphorylation drastically alters steric, chemical and electrostatic properties of proteins, changing protein compactness and introducing new possibilities for intra and intermolecular electrostatic interactions (Mandell et al., 2007). Also acetylation causes changes of protein charges and hence it shares with phosphorylation some conformational effects (Mao et al., 2010; Marsh and Forman-Kay, 2010).

Different kinds of multiple PTMs can occur on a same IDP, giving rise to combined and rather complex effects. One illustrative example is given by histones: they receive methylation, acetylation, phosphorylation, ubiquitylation, ADP-ribosylation, and SUMOylation. These PTMs occur at different stages of their action, affecting histone-histone and histone-DNA interactions and thus

influencing the nucleosome stability (Liu et al., 2012).

1.1.5 The conformation energy landscape of IDPs

Drawing the conformation energy landscape of IDPs may help to understand the "diversity" of IDPs with respect to globular proteins. An IDP is poorly represented by a single, lowest-energy conformation and can be better defined as a dynamic ensemble of interconverting conformers. The fluctuating, dynamic equilibrium among iso-energetic minima can be described by a shallow energy landscape (Flock et al., 2014). Figure 5 compares typical energy landscape of IDPs, well-folded proteins and complexes resulting from the interaction between them. As expected, the profile of globular proteins exhibits a single global energy minimum, which corresponds to the native state (**Figure 1.5 a/d**), while IDPs have a high energy profile (**Figure 1.5 b/e**). Interaction between an IDP and its binding partner may give rise to a new energy profile (**Figure 1.5 c/f**). Many IDP conformations still "fluctuate" at high energy, while a part of the conformational ensemble is "frozen" in the bound form and reaches a minimum through a quite narrow stem. This phenomenon is also referred as "folding upon binding". Such a transition may sometimes involve just a region of an IDP, while other regions remain disordered, giving rise to a so-called "*fuzzy complex*" (Tompa and Fuxreiter, 2008). Changes of conformational free energies may be induced also by environmental conditions, PTMs, and interactions with other (macro)molecules (Boehr et al., 2009; Flock et al., 2014; Fuxreiter, 2012; Kar et al., 2010; Ma and Nussinov, 2009).

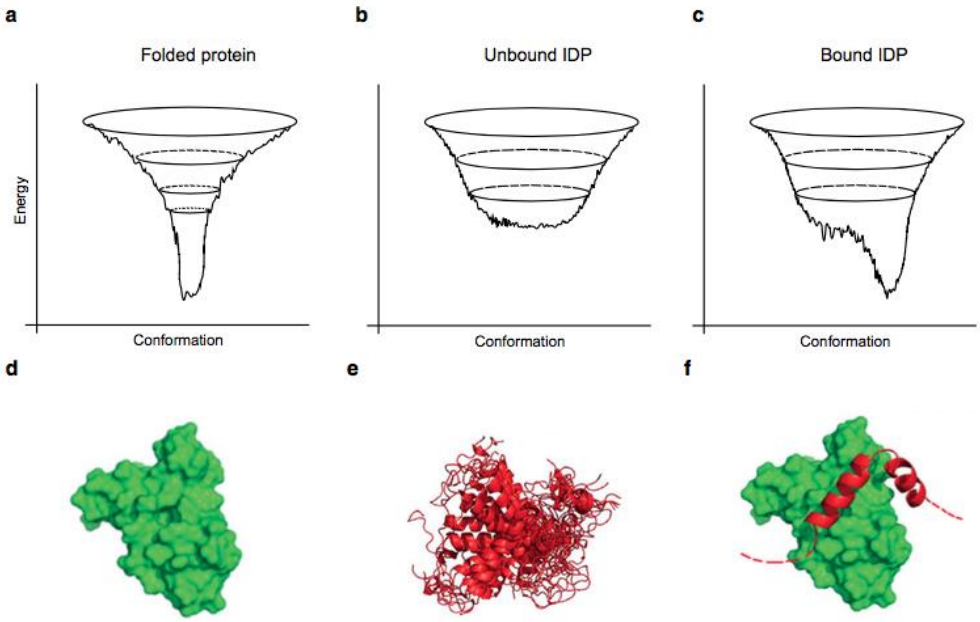


Figure 1.5. Representations of typical conformation energy landscapes and structure of folded and disordered proteins, and their complex. Folding funnel for a) a well-folded protein in which the global minimum corresponds to the native state; b) an IDP characterised by several close energetic minima representing the lowest energy conformational states; c) an IDP underwent to binding-induced folding. Note the consequent modification of the energy landscape and stabilization of a single conformation. Panels d-f) show the corresponding schematic 3D structures (Pauwels et al., 2017).

1.2. Biological relevance of structural disorder

This chapter deals with the abundance of IDPs in different proteomes, their involvement in physiological and pathological events.

According to the 3-D structures to date retrievable in PDB, only a minority of proteins (ca. 32%) can be considered "disorder-free" (Dunker et al., 2013; Uversky, 2013a). This amount may be overestimated, since disorder is elusive to high-resolution techniques devoted to structural studies and fully-disordered proteins are poorly represented in PDB. Many data on the occurrence of structural disorder come from computational analyses of amino acid data banks.

1.2.1 Occurrence of IDPs in proteomes

The natural abundance of IDPs/IDRs has been neglected until the first bioinformatic systematic investigations have been undertaken on proteome databases. This work has indicated that about 25-30% of eukaryotic proteins are mostly disordered (Uversky et al., 2005), and that more than half part of mammalian proteins has long (>30 residues) regions of disorder (Dunker et al., 2000). More in detail, long disordered segments were found to occur in 2.0% of archaea, 4.2% of eubacterial and 33% of eukaryotic proteins (Ward et al., 2004). These data highlight the high frequency of IDPs in more complex organisms. A more recent and extensive study has considered around 3500 proteomes from viruses and three kingdoms of life (Xue et al., 2012). This work essentially confirms the previous observations, although it has not taken into account single proteins, but the average fraction of disordered residues in the analysed proteomes (Xue et al., 2012). Figure 6 shows the fraction of disordered residues with respect to the proteome size (Xue et al., 2012). A general observation is that among archaea, prokaryotes and eukaryotes, disorder increases with proteome size. The viruses represent an exception that deserves to be considered apart (**Figure 1.6**).

A well-defined gap exists between the frequency of disordered residues in

prokaryotes (≤27%) and eukaryotes (≥32%) (Xue et al., 2012). Moreover, the highest eukaryotes exhibit more disorder than unicellular one. This suggests that intrinsic disorder increases with organism complexity. An exception is represented by a small group of highly "disordered" unicellular eukaryotes, which are parasitic host-changing protozoa. The wide variability of their habitats during their life-span might have required a complex equipment of metabolic answers, which can be obtained through the increase of structurally and functionally promiscuous proteins.

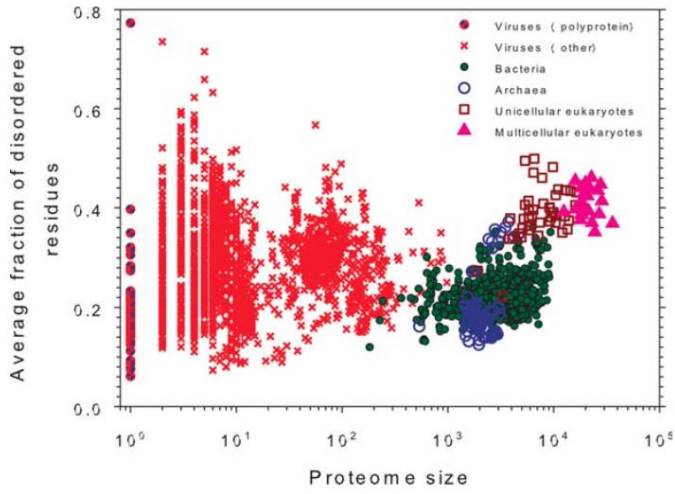


Figure 1.6. Correlation between the content of intrinsic disorder and the proteome size for 3484 species of viruses, archaea, bacteria, and eukaryotes. Each symbol indicates a species: small red circles filled with blue indicate viruses, small red circles indicate other viruses, small green circles indicate bacteria, blue circles indicate archaea, brown squares indicate unicellular prokaryotes and pink tringles indicate multicellular eukaryotes (Uversky, 2013a).

As already anticipated, it is striking the high frequency of disorder associated to virus proteomes. Indeed, it has been hypothesized that the characteristics of IDP/IDRs are well suited to extremely "compact" proteomes as those of RNA viruses, where a limited number of multi-functional proteins might accomplish variegated roles (Stamm et al., 2005; Uversky and Dunker, 2010). The flexibility and chameleonic features of IDRs may reveal useful also to cope with the host immune system. Moreover, disordered regions could "buffer" the deleterious effects of mutations introduced by low-fidelity viral polymerases better than structured domains would. Even for viral proteomes, it can be hypothesized that the exposition to highly variable environment plays a crucial role in inducing higher structural disorder (Xue et al., 2012; Xue et al., 2010b).

1.2.2 Biological roles of IDPs as interaction hubs

The abundance of structural disorder among proteomes suggests it can be associated to important biological functions (Uversky, 2011). As previously observed for virus proteomes, the multitasking activities of IDPs/IDRs represent one of the solutions used by Nature to increase the organism complexity without expanding the genome size. The ability of IDPs/IDRs to fulfil more than one function makes them to belong to a special class of "moonlighting proteins" (Jeffery, 2003, 2004; Tompa et al., 2005), and consequently to be promiscuous. Often, IDPs serve as hubs or "scaffolds" in protein interaction networks (Dyson and Wright, 2005). A scaffold protein is placed at the centre of functional complexes, where it interacts with most of its partners at the same time. Typically, the architecture of scaffold proteins includes several small globular domains (~80 amino acids, on average) connected by long linker regions (~150 residues, on average) with crucial binding functions (Balázs et al., 2009).

The mechanism of binding is often mediated by short recognition elements or motifs (Fuxreiter et al., 2007; Neduva et al., 2005). They have been classified

mainly based on their length (3 to 30 residues), the existence of preformed structural elements (PSEs) in the free state, and the persistence of disorder upon the formation of a complex (Chen et al., 2006; Kim et al., 2003; Tompa, 2012b).
Among the regions mediating the interactions are the so-called MOlecular Recognition Features (MORFs). They contain 20-30 residues that typically undergo a disorder-to-order transition stabilized by binding to a partner. A MoRF can be further classified according to the structure can adopt in the bound state. Indeed, there are α-MoRFs, β-MoRFs, and ι-MoRFs which form α-helices, β-strands, and irregular (but rigid) secondary structures, respectively (Mohan et al., 2006). Moreover "complex MORFs" contain combination of several types of secondary structure. However, comparison of free and bound structures, experimentally observed or predicted, suggests that IDPs have rather strong preferences to reach α-helical conformations (Liu et al., 2006).
Linear Motifs (LM) (Chen et al., 2006; Davey et al., 2012; Diella et al., 2008) are short sequence motifs (3-10 amino acids long) within a more ordered environment (Tompa, 2012b). They are enriched in hydrophobic residues (Trp, Leu, Cys, and Tyr), charged (Arg and Asp), and Pro residues, and they are depleted in Gly and Ala. LMs are poorly specific in terms of primary structure (Forman-Kay and Mittag, 2013) and highly flexible, which allows them to adopt various conformations and to bind to multiple partners. LMs show several functions: they target proteins to a subcellular location, recruit PTM enzymes or binding factors, thus controlling the protein stability and the formation of complexes (Davey et al., 2012; Diella et al., 2008).
Mis-identification of binding partners and mis-signalling represent "loss-of-function" events resulting in a number of pathologies, due the involvement of IDPs as interaction hubs in many crucial biological processes. Beside this, the pathological role of IDPs is related to their misfolding, which leads to aggregation and/or fibril formation. This is the topic of following paragraph.

1.2.3 Pathological effects of IDP aggregation

Proteins may misfold giving rise to amorphous, native-like or fibrillar, highly-ordered aggregates (Dobson and Chiti, 2017). *In-vivo* aggregation implies "gain of function", which has detrimental biological consequences in most documented cases. Indeed, massive fibrillar aggregations of some IDPs cause severe cellular/tissue/organ damages that are related to well-known amyloid pathologies, such as diabetes, Parkinson's, Alzheimer's, and cardiovascular diseases (Uversky et al., 2008). Therefore, most studies on IDP aggregation deals with fibrillization. Fibrils contain predominantly β-sheet structure in a typical cross-β conformation, independently of the primary structure of involved protein. Structural studies indicate that fibrillation precursors, or "protofilaments", are composed by a variable number of protein monomers, which assemble in β-sheet conformation. The β-strands are perpendicular to the fiber axis, held together by hydrogen bonds involving side chains and running parallel to the fiber axis. Individual "protofilaments" are often twisted one around another to form long, straight and unbranched mature fibrils (Rambaran and Serpell, 2008).

When amyloids are formed from a globular protein, an unfolded intermediate may be required, which exposes hydrophobic residues to promote intermolecular interactions (Kim and Hecht, 2006). On the contrary, fibrillization of IDPs may require folded or partially folded intermediates which act as polymerization "seeds" into amyloid fibrils (Uversky and Fink, 2004) (**Figure 1.7**). So, what induces "ordering" and aggregation in IDPs?

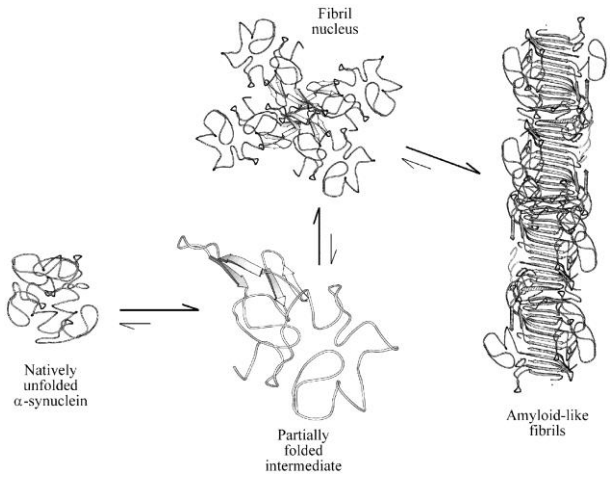


Figure 1.7. Aggregation pathways of α- synuclein (Uversky *et al.*, 2001)

Aggregation determinants may be intrinsic or extrinsic. Intrinsic factors are related to protein sequence, while extrinsic factors are more complex and include macromolecular crowding inside the cell/cellular compartments, contact with membrane lipids, environment pH, presence of chaperones, small molecules or metal ions and post translational modifications (Breydo et al., 2017). I will not consider exhaustively all these factors and I will focus on some of them.

Overall, the relevance of sequence determinants is witnessed by the number of amyloid diseases caused by mutations (Chiti et al., 2002). More in detail, an interesting hint comes from the analysis of natural mutations modifying the net charge of the proteins or protein fragments associated with familial forms of amyloid diseases. It emerges that reduction of the net protein charge is an important determinant in some amyloid diseases (Chiti et al., 2002). *Viceversa*, the importance of charge to avoid protein aggregation has been recognized not only for evolutionary reasons, but also for its medical (Prabakaran et al., 2017; (Sant'Anna et al., 2014) and biotechnological implications (De Baets et al., 2015; Prabakaran et al., 2017).

It must be observed that in the case of Huntington's disease and some other "polyglutamine diseases" the causative mutation is the expansion of Gln repeats

(poly-Q tract) (Edbauer and Haass, 2016). Polyglutamine-expanded proteins are more prone to cleavage and fragments containing the Poly-Q tracts have a higher propensity to undergo misfolding and aggregation. This example pinpoints that also polar, or non-hydrophobic, residues can be involved in aggregation events.

Among extrinsic factors, the contact with membrane lipids has been recently explored with different approaches by several research groups. Amyloidogenesis has been hypothesized to occur by membrane-mediated mechanisms, especially for neurodegenerative diseases, such as Alzheimer's disease. Indeed, the brain represents a lipid-rich environment and oxidative damage of lipids has been often correlated with aggregation of amyloid-β (Aβ) causing Alzheimer's disease. A possible chemical mechanism linking oxidative stress with amyloid formation involves an oxidative by-product of unsaturated lipid, the 4-hydroxy-2-nonenal (HNE). It has been demonstrated that Aβ has itself a pro-oxidant activity which promotes the production of HNE. The interaction of Aβ with HNE can in its turn modify Aβ, increasing its affinity for lipid membranes and its tendency to aggregate into amyloid fibrils (Murray et al., 2007). More recent studies indicate that oligomers of Aβ may disrupt the bilayer integrity and, *viceversa*, can be modified by lipids. Indeed, the interaction with lipids can cause the fragmentation of preformed fibrils into remodelled, toxic protofibrils, which speed up the aggregation through secondary nucleation steps (Korshavn et al., 2017; Lindberg et al., 2017).

Among extrinsic factors leading to aggregation, the most relevant to the topic of this thesis is pH. It has been demonstrated that formation of Ribonuclease Sa fibrils can be experimentally induced by shifting environment pH to the protein pI (Schmittschmitt and Scholtz, 2003).

The issue of IDP solubility/aggregation represents an important issue, considering not only fibrillization but also the physiological role of condensation in *spatio*-temporal organization of cellular functions.

1.2.4 Role of IDPs in cellular phase transition

The conversion of a highly dynamic ensemble of conformers into less disordered aggregates can concern several IDPs, although they are normally highly soluble and contain a few hydrophobic residues (Chiti and Dobson, 2006). This aspect has been initially explored because of the involvement of IDPs in the formation of highly ordered amyloid fibrils (Uversky et al., 2001). More recently, it has emerged that IDPs are involved in very important physiological phenomena, such as protein condensation, or "collapse", giving rise to membrane-less organelles, such as nucleoli or Kajal bodies, stress granules etc (Wu and Fuxreiter, 2016). The condensation is referred to the crowd of heterogeneous mixtures of proteins and nucleic acids, bringing to a phenomenon similar to polymer condensation, in response to various metabolic and stress stimuli. Membrane-less organelles allow a dynamic cell compartmentation, and, hence, *spatio*-temporal control of biological reactions (Wu and Fuxreiter, 2016). It has been observed that almost invariably condensation of membrane-less organelles involves IDPs (Uversky, 2017).

Which is the relevant feature of IDPs in this context? It seems that the emerging property of IDPs is not related to their high solubility, or high propensity to aggregate, but to their promptness to conformational changes, which results from a fine interplay among backbone, sidechains and solvent-mediated interactions. This property is very likely to be dictated by the sequence (Holehouse and Pappu, 2018). Let's consider the effect of different class of amino acids. Although not frequent in IDPs, hydrophobic residues can lead condensation through the formation of expanded clusters, as witnessed by recent studies on P domain of Pab1 (Riback el al., 2017). Polar residues may drive condensation through the formation of intramolecular hydrogen bonds, dipole-dipole interactions, or either amide-amide hydrogen bonds, which can be entropically favored with respect to amide-solvent ones upon protein collapse (Holehouse and Pappu, 2018). An

example of gelation and phase separation driven by polar residues is given by the Glu/Asp-rich domain of yeast prion Sup35 (Molliex et al., 2015)

What about charged residues? We have already seen that highly charged IDPs, endowed with high values of NCPR can be highly expanded, coil-like ensembles (Mao et al, 2010). However, charge interactions can also drive compaction, according to the patterning of charged residues and the formation of intramolecular attractive interactions (Das and Pappu, 2013). Similarly, attractive intermolecular interactions may drive the condensation. It is likely that such a network of interactions is modulated by PMTs directly affecting the protein charges (i.e. phosphorylation or acetylation), or by changes of pH and temperature (Holehouse and Pappu, 2018). Indeed, the entropic cost of solvating charged groups increases with temperature, thus favoring intra/inter-chain interactions (Wuttke et al., 2014).

Other features may favor the involvement of IDPs in phase-transition phenomena. Not only IDPs can give a fast and concerted response to environment stimuli, but their response is most often reversible, due to the lack of structural elements and of complex hierarchical organization. Moreover, IDPs are able to detect and to intensely respond to even subtle signs from the environment. Here the low-complexity of their sequence can play a key role, by "amplifying" the compositional elements acting as "*antennae*" and "*effectors*".

2. Methods

2.1 Computational and Experimental techniques used in this work

The study of IDPs requires specific methods and techniques imposed by their structural and biophysical properties which are peculiar with respect to globular proteins. Usually, several techniques, based on independent physical or chemical principles are used in combination, to obtain complementary results.

This chapter does not describe in detail the experimental procedures also referred in the section entitled "Experimental work". Instead, it illustrates theoretical aspects of used bioinformatics, biochemical and biophysical methods.

2.1.1 Computational techniques

The development of various disorder predictors has been mainly based on sequence analysis and has revealed a useful tool for large-scale proteomic investigations. The repertoire of bioinformatics tools is nowadays rather wide and based on different concepts, physicochemical parameters and implementation techniques. It is difficult to establish the best predictor and it can be useful to combine some of them to obtain a more reliable result.

Here, we show three different tools based on sequence analysis and applied in this work: the charge-hydropathy plot (CH plot) (Uversky, 2002b), the meta-predictor Pondr-fit (Xue et al., 2010a); and CIDER (Holehouse et al., 2017).

Charge-Hydropathy plot (CH)

A rather simple approach to predict the intrinsic disorder of a protein is based on the empirical observation that ordered and disordered proteins exhibit different average net charge and hydropathy (Uversky, 2002b). The CH plot compares the absolute, mean net charge - neglecting histidine - and the mean, scaled Kyte-Doolittle hydropathy (**Figure 2.1**). The hydropathy is scaled between 0 and 1. Ordered and disordered proteins plotted in this charge-hydropathy graph can be separated by a linear boundary. The output of this simple tool is binary, but the

distance from the separating line may carry information on the extent and type of disordered on the whole chains. This method gives an estimated overall classification accuracy of 83% with 76% for disordered proteins and 91% for ordered proteins. A limitation of CH plot is that it allows only a binary classification of proteins, without providing information at amino acid resolution.

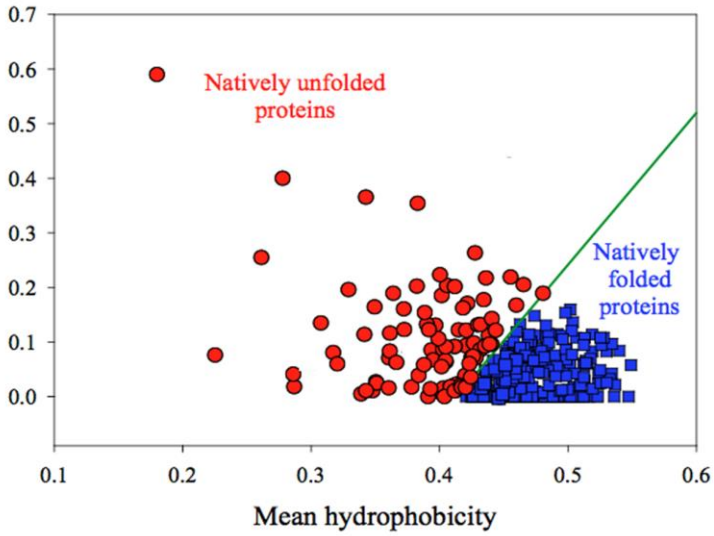


Figure 2.1. CH plot presents mean net charge versus mean hydropathy plot for sets of 275 folded proteins (blue squares) and 91 IDPs (red circles) (Uversky *et al*., 2000).

Pondr-fit, a meta-predictor of intrinsically disordered amino acids

Pondr-fit is a meta-predictor of structural disorder, it means that the final result comes from a collection of predictors used as inputs for another one. In this way, it is possible to obtain an improvement in accuracy because different predictors extract information from different sequence features, prediction models, and training sets. All of them use the primary sequence as the input and give an individual score output, one for each amino acid in the sequence, indicating each residue's likelihood of being structured or disordered (**Figure 2.2**). The individual predictors used in the analysis are PONDR VLXT (Romero et al., 2001), PONDR VL3 (Peng et al., 2006), PONDR VSL2 (Peng et al., 2006), IUPred (Dosztányi et

al., 2005), FoldIndex (Prilusky et al., 2005), and TopIDP (Campen et al., 2008). Pondr-fit, was found to improve the prediction accuracy over a range of 3 to 20% with an average of 11% compared to the single predictors, depending on the datasets being used. Analysis of the errors shows that the worst accuracy still occurs for short disordered regions with less than ten residues, as well as for the residues close to order/disorder boundaries. The understanding of the mechanisms by which such meta-predictors may improve their predictions will likely promote the further development of protein disorder predictors.

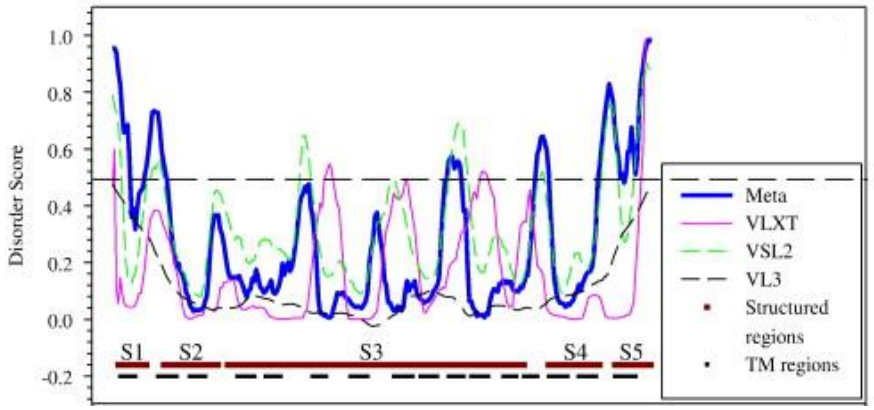


Figure 2.2. Prediction of intrinsically disordered residues in Human P53 by Pondr-fit and its 6 predictors. Brown lines S1-S5 are structured DNA-binding domain. The dashed line at 0.5 of Y-axis is a threshold for disordered/structured residues. Residues with a score above this line are predicted disordered, and residues with a score below 0.5 are predicted to be ordered. Meta, VLXT, VSL2, and VL3 correspond to the prediction from Pondr-fit, PONDR VLXT, PONDR VSL2, PONDR VL3, respectively (Uversky and Dunker, 2010).

Classification of Intrinsically Disordered Ensemble Regions (CIDER)

CIDER is a webserver developed by the Pappu lab, that allows to assign a conformational class to a sequence and calculate a number of key parameters from the primary sequence of IDPs concerning charge distribution along the sequence (Das and Pappu, 2013; Das et al., 2015). Using the fractions of positively charged (*f*+) and negatively (*f*-) charged residues (FCR), IDP sequences can be partitioned into one of five conformational classes (from R1 to R5) in a diagram of state (**Figure 2.3**). This diagram has been used to annotate IDP sequences from DisProt databse (Sickmeier et al., 2006) filtered for low overall hydrophobicity and low overall proline content (<15%). Region 1 represents low-FCR sequences that adopt globular conformations, region 2 contains a variety of conformations, from the compact globules to the swollen coils, while region 3 accommodates high-FCR sequences, which are mainly polyampholytes in non-globular conformations (*i.e.* coil-like and hairpin-like) (Das et al., 2015). Noteworthy, in region 3 and upper region 2, the linear distribution of opposite electrostatic charges along protein sequences seems to determine the conformational compactness. Region 4 and region 5 house respectively completely negative and positive proteins.

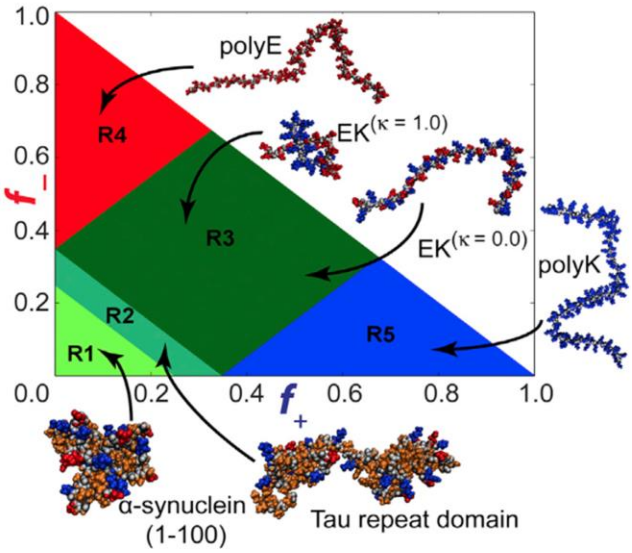


Figure 2.3. Diagram of state annotated with representative conformations for specific IDPs that correspond to each of the five regions (Holehouse *et al*., 2017).

2.1.2 Biochemical techniques to experimentally assess structural disorder

Often SDS-PAGE can be used to assess structural disorder. Indeed, IDPs and IDRs exhibit a lower mobility compared to that of equally-sized globular proteins and an apparent molecular weight (MW) 1.2–1.8 times higher than the real one. This behaviour has been ascribed to lower ability to bind sodium dodecyl sulphate (Tompa, 2002) and, hence, reflects the amino acid composition of IDP/IDRs.

The protease sensitivity assay is one of the earliest methods set up to recognize structural disorder (Hipp et al., 1952). IDPs and IDRs are much more sensitive to the activity of proteolytic enzymes than globular proteins (Morin et al., 2006; Tompa, 2002). From IDP composition also stems their atypical response to environment (temperature, pH, molecular crowding, strong denaturants) (Uversky and Longhi, 2011).

IDPs revealed rather insensitive to denaturing conditions, including heating. It has been even reported a partial, reversible folding, through the formation of secondary structure in response to heating. This phenomenon has been described

for α-synuclein (Uversky et al., 2001), the extracellular domain of nerve growth factor (Timm et al., 1994) and a_S-casein (Kim et al., 2000). The general heat stability of IDPs allows their purification by incubating row cell extracts or cell suspensions at high temperature, which selectively and irreversibly denature globular proteins. This procedure has been firstly reported for dehydrins, a class of vegetal IDPs recombinantly expressed in *E. coli* (Livernois et al., 2009).

2.1.3 Biophysical techniques to experimentally asses structural disorder

IDPs often exist as a dynamic ensemble of conformations, so this precludes in most cases the application of high-resolution techniques able to solve the 3D structure of the proteins, such as X-ray crystallography and nuclear magnetic resonance. Hence, structural characterization of IDPs is possible through combined, complementary physicochemical approaches. Some of them are briefly described below.

Determination of Stokes' or hydrodynamic radius

An insightful parameter to define protein conformation is hydrodynamic or Stokes' radius (R_S). The R_S is defined as the radius of a hard sphere that diffuses at the same rate as that of solute (**Figure 2.4)**. The R_S can be estimated, under the Stokes' law assumption (a perfect sphere traveling through a viscous liquid), through the following equation:

$$R_{S=} \frac{T k_B}{6 \pi \eta D}$$

where: k_B is the Boltzmann constant, T is the temperature, η the medium viscosity, D the diffusion constant.

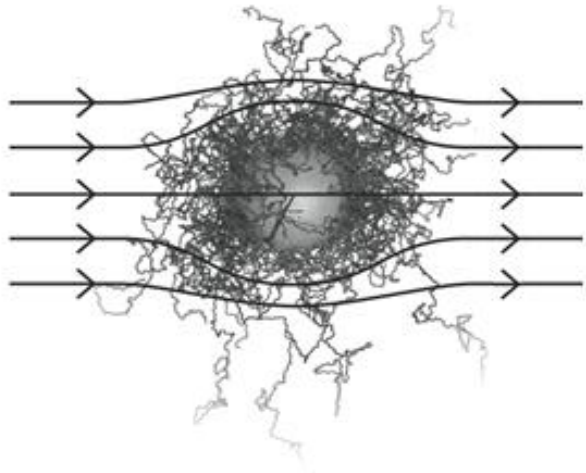


Figure 2.4. Visual representation of an IDP as a sphere moving in a fluid (Nygaard *et al*., 2017).

The R_S can be measured using size exclusion chromatography (SEC) or dynamic light scattering (DLS).

DLS is a technique based on measuring the random changes in the intensity of light scattered from a suspension or solution. It's an appropriate technique to monitor the size of protein molecules (Uversky and Longhi, 2011).

SEC consists in a separation technique where the stationary phase is composed by porous beads. Molecules suspended in the mobile phase can pass through the resin at different rates according to their R_S.

The calibration curve correlating elution time/volume to R_S are usually obtained with globular and well-known proteins. The plot representing the ratio Ve/Vo against the log of standard hydrodynamic radii is described by a linear function, which can be used to estimate the R_S once known the Ve/Vo of a given protein. IDPs appear endowed with a larger R_S if compared to equally-sized native globular proteins (**Figure 2.5**).

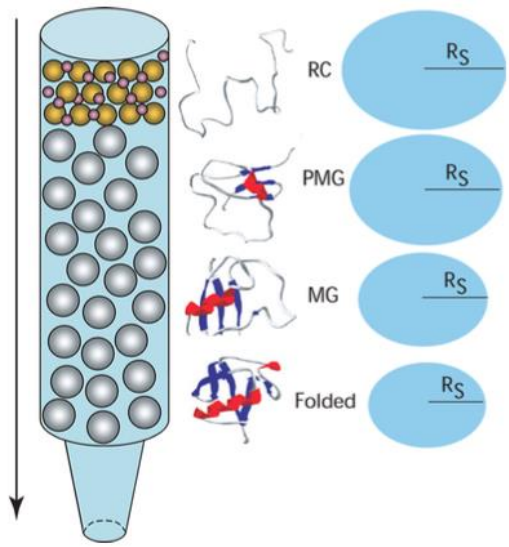


Figure 2.5. Size exclusion chromatography (SEC) analysis of conformational behaviour of proteins. Schematic representation of the physical principles of molecule separation by SEC. The stationary phase is made of porous beads represented as grey spheres. The sample applied on the top of the column contains small and large molecules represented as pink and yellow spheres, respectively. A vertical arrow on the left indicates the flow direction of mobile phase. On the right, relative hydrodynamic volumes occupied by the same polypeptide chain in four different conformations: RC, random coil; PMG, pre-molten globule; MG, molten globule; and folded (Habchi et al., 2014).

Hence, the behaviour of IDPs is described by different sets of empirical equations suitable for different conformational categories, such as globular or natively folded (NF), pre-molten globule (PMG), and random coil (RC) proteins. Moreover, it has been defined the relationship between Rs and molecular mass (MM) or sequence length for each of these categories (Uversky, 1993, 2012; Wilkins et al., 1999).

$$\text{Log } (R_S^{NF}) = 0.369 \text{ Log } (MM) - 0.254$$

$$\text{Log } (R_S^{RC}) = 0.521 \text{ Log } (MM) - 0.649$$

$$\text{Log}\,(R_S^{PMG}) = 0.403\ \text{Log}\,(MM) - 0.239$$

The R_S of an IDP with N residues can be also calculated according to (Marsh and Forman-Kay, 2010), using the simple power-law model:

$$R_S^{IDP} = R_0 N^{\nu}$$

where $R_0 = 2.49$ and $\nu = 0.509$.

Similar equation where formulated to calculate the R_S as exponential function of the amino acids number (N) (Wilkins et al., 1999). Note that the Wilkins' empirical equations can be applied only to globular and unfolded proteins:

$$R_{S_{(globular)}} = (4.75 +/-1.11)\ N^{\,0.29 +/-0.02}$$

$$R_{S(unfolded)} = (2.21 +/- 1.07)\ N^{\,0.57+/- 0.02}$$

Overall the conformation of any given IDP (coil-like, PMG-like, or molten globule-like) can easily be discriminated by its R_S and the ratio between the experimental and the theoretical value expected for a globular protein of equal size.

Determination of gyration radius

Another useful parameter for defining protein conformation is gyration radius (R_g). The R_g of a solid sphere can be indicated as a point at the distance $r\sqrt{3/5}$ from its mass centre, where *r* is the sphere radius (**Figure 2.6**). In the case of IDPs, the definition of R_g is conceptually borrowed from the polymer physics and is used to represent the dimension of a polymer chain whose conformations change with time, reaching a *quasi*-infinite number. Hence, in the field of polymer physics, the "radius of gyration" is intended as a mean over all polymer molecules of the sample and over time. For an ideal polymer, R_g can be considered proportional to the mean squared end-to-end distance over all polymer molecules of the sample and over time, R_e (Flory and Volkenstein, 1969).

Indeed, $Rg^2 = \langle d^2 \rangle / 6$, being d the end-to-end distance, with angular brackets indicating the average over all the configurations. In heteropolymers, such as proteins, the "decoupling" of R_g and R_e it is likely to occur because the chemical heterogeneity of interactions (Holehouse and Pappu, 2018).

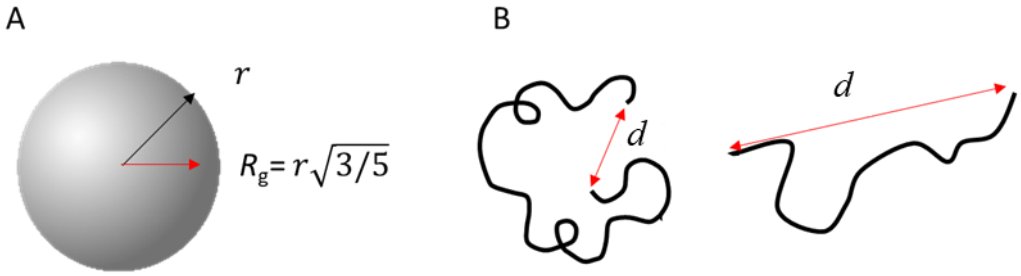


Figure 2.6. Visual representation of Rg for a sphere (A) and end-to-end distance in a polymer (B). A) Rg has an invariable value proportional to the sphere radius *r*. B) Two "extreme" conformations (condensed and extended) of the same polymer show different end-to-end distances represented by *d*.

Measurements of R_g can be obtained through small angle X-ray scattering (SAXS), a powerful method for structural characterization (sizes and shapes) of disordered and ordered proteins. Setup of SAXS is conceptually simple: a solution of proteins usually placed in quartz capillary is illuminated by a collimated monochromatic X-ray beam; the intensity of scattered X-rays is recorded by an X-ray detector. Sample scattering intensity is proportional to concentration and protein dimension. SAXS provides low resolution structural data and gives access to the mean particle size (R_g) as well as to the maximal intramolecular distance (D_{max}), which are related with the degree of compaction/extension of the molecule. R_g is smaller for proteins with a compact shape as compared to extended proteins with identical amino acids. The structural properties of the polypeptide chain can also be determined by the R_S/R_g ratio. This ratio should be $(3/5)^{1/2}$ for a globular

protein, around 0.9 for a pre-molten globule, and > 1.5 for a random coil (Gast et al., 1994). The SAXS data on their own do provide several indicators of the presence of protein flexibility:

In some case it is more intuitive to interpret the structural properties analysing the pair distribution plot P(r) rather than the scattering itself. The P(r) function can be obtained from the experimental scattering data using indirect Fourier transformation (Glatter, 1977; Svergun, 1992). Globular compact particles have a symmetric bell-shaped P(r), whereas unfolded particles have an extended tail (**Figure 2.7 B**). A very useful method to discriminate between different structural conformation is the Kratky plot (**Figure 2.7 C**): for a globular protein it has a typical bell shape with a clear maximum, for a completely unfolded protein or in a pre-molten globule conformation, no such maximum can be observed, and the curve displays a plateau (Glatter and Kratky, 1982).

Highly-flexible proteins, such as IDPs, can exhibit conformational states differently populated. Therefore, their accurate structural description requires its definition in term of ensemble: EOM (Ensemble Optimization Method) is a very useful algorithm for characterising them.

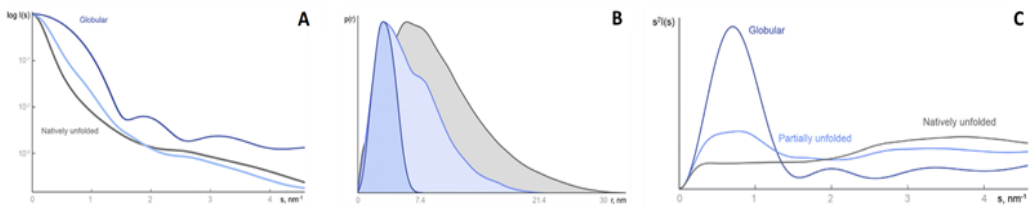


Figure 2.7. Data simulated from three 60 kDa proteins: globular (dark blue), 50% unfolded (light blue) and fully disordered. A) Logarithmic plot, B) Distance distribution functions P (r), C) Kratky Plot (Kikhney and Svergun, 2015).

Secondary structure of IDPs: spectroscopic techniques

The content of secondary structure is another aspect that can be useful to describe IDP conformation. In this study, secondary structure contribution has been analysed through circular dichroism (CD) and Fourier transform infrared (FT-IR) spectroscopy.

CD spectroscopy employs left- and right-handed circularly polarized light, which is differentially absorbed by secondary structures in a protein (Atkins and De Paula, 2013). In the far UV (180 - 250 nm), the CD of a protein is primarily that of the amide chromophores along the backbone, which result from bonding between the component amino acids. α–helix, β–sheets and unordered polypeptides show different specific peak (**Figure 2.8**). α–helix structure is characterized by a maximum at 192 nm and two minima of similar magnitude at 208 and 222 nm. The structure of β–sheet shows a single minimum at around 216 nm and a positive peak of comparable magnitude near 195 nm. Unordered or random coil peptides present a deep minimum of ellipticity signal just below 200 nm (Nordén, 1997).

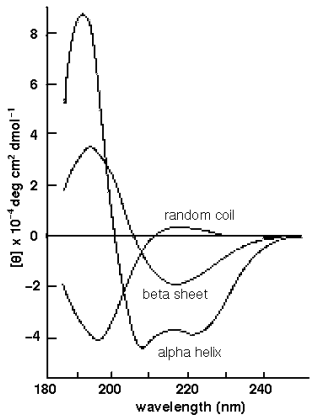


Figure 2.8. CD spectra of α–helix, β–sheets and unordered polypeptides. (http://www.cryst.bbk.ac.uk/PPS2/assignments/A1/CD_info.html).

FT-IR is a powerful method to investigate protein aggregation, secondary structure and stability through the analysis of the protein absorption spectrum. Absorption in the infrared region results in changes in vibrational and rotational status of the molecules: absorption frequency depends on the vibrational frequency of the molecules; whereas the absorption intensity depends on how effectively the infrared photon energy can be transferred to the molecule, and this depends on the change in the dipole moment that occurs as a result of molecular vibration. Thus, all compounds (except elemental di-atomic gases such as N_2, H_2 and O_2) have typical infrared spectra and most components can be analysed by their typical infrared absorption. The main absorption bands that have been widely used for protein characterization are the so-called amide I (1700–1600 cm^{-1}), amide II (1600–1500 cm^{-1}), and amide III (1400–1200 cm^{-1}), which are due to the vibrational modes of the protein backbone (**Figure 2.9**). Band in amide I is due to the stretching vibration of the C=O peptide bond, is very sensitive to the C=O environment and consequently to the secondary structure of the proteins. Band in amide II is due to the contribution of several backbone modes, —NH in-plain bending and CN stretching, with small contributions from C=O bending, CC and NC stretching vibrations. It is also sensitive to the protein secondary structure, but its analysis is complicated by the overlapping of the different vibrational modes.

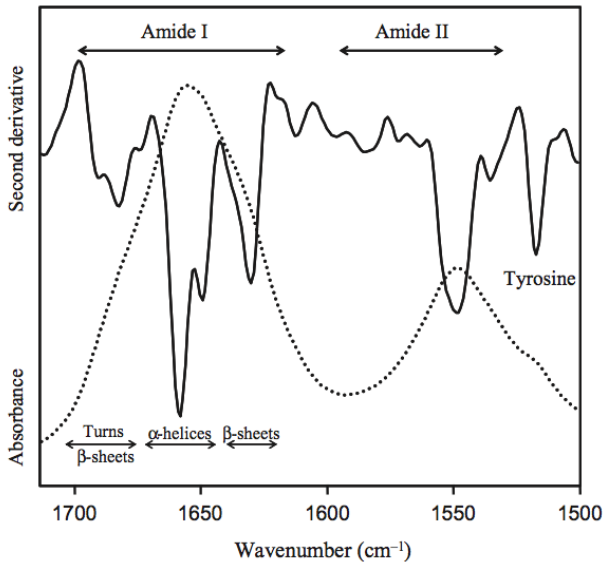


Figure 2.9. Absorption spectrum of CRL1 in water (dotted line), and its second derivative in the amide I and amide II regions (continuous line). The tyrosine band around 1515 cm^{-1} is indicated (Natalello and Doglia, 2010).

As regard secondary structure, the absorption of helices **(Figure 2.9**) occurs in water in the region 1660–1648 cm^{-1}. The band position depends on α–helix length, flexibility, and hydration. In particular, higher wavenumbers characterize short and flexible helices, while lower wavenumbers are associated to long and rigid structures (Arrondo et al., 1993). Intramolecular β-sheets display two absorption bands of different intensity (1633 cm^{-1} and 1686 cm^{-1}) (Bath and Zscherp, 2002). Intermolecular β-sheets display an absorption profile similar to that of native intramolecular β-sheets, but shifted in peak positions. In amyloid and thermal aggregates, as well as in bacterial inclusion bodies, the low-frequency band was found to be in the range 1630–1620 in water, while the high-frequency band was observed between 1698–1692 cm^{-1} in water. The absorption of the C=O group in random coil structures is due to the contributions of the peptide bonds in different environments, therefore leading to a broad band centred around 1654 cm^{-1} in

water. Unfortunately, this band is superimposed to the α-helix secondary structure. The absorption of other secondary structures occurs in the wide amide I range (Arrondo and Goñi, 1999; Bath and Zscherp, 2002; Goormaghtigh et al., 1994; Susi and Byler, 1986; Tamm and Tatulian, 1997). Among them, β-turns can be found from 1686 to 1660 cm^{-1} in water (Goormaghtigh et al., 1994).

In conclusion, in some case, the use of D_2O instead of water allows to discriminate better secondary structure (**Table 2.1)**.

Secondary Structure	Band Position (cm^{-1})			
	H_2O		D_2O	
	Average	Extremes	Average	Extremes
Turn	1672	1686–1662	1671	1691–1653
α-Helix	1654	1660–1648	1652	1660–1642
Random coil	1654	1657–1642	1645	1654–1639
β-Sheet	1633	1640–1623	1630	1638–1615
intramolecular	1686	1695–1674	1679	1694–1672
β-Sheet	1625	1630–1620	1620	1630–1611
intermolecular*	1695	1698–1692	1686	1690–1680

Table 2.1. Average band positions and spectral ranges in H_20 and D_2O (Natalello, 2010).

3. Aims, main results and conclusions

Charged residues may have an upmost role in controlling two aspects of IDP life, which are aggregation/solubility and compactness.

The issue of how charges can induce IDP aggregation has been addressed considering also the effects of environment *stimuli* and namely of pH changes. More in detail, we have studied the solubility of a set of model disordered proteins, namely PNT variants, differing in their net charge per residue, NCPR, but sharing the same hydrophobicity. The starting point is the concept that an IDP, similarly to a globular one, may respond to pH changes as described by its titration curve. The hypothesis leading the experimental work considers that the promptness or intensity of its response to pH changes must depends not simply on the net charge, which dictates the pI, but on its NCPR, which commeasures the net charge to the protein size.

The chosen protein for this first work is the well-studied PNT. It behaves as a molten globule (Habchi and Longhi, 2012), can be produced at high level as recombinant production and it is easily purified by affinity chromatography. Hence, starting from the native sequence of PNT, a set of mutants have been rationally designed aimed at sampling different values of NCPR, but sharing the same hydrophobicity. Two mutants were obtained by drastically changing the number of basic and acidic residues present in the *wt* protein. Hence, the whole set of PNT variants includes proteins with a strongly acidic, a mild acidic and strongly basic pI. Variants of PNT were recombinantly produced and their behaviors compared with that of a single globular protein, the green fluorescence protein (GFP), which shares very similar hydrophobicity, length and features a mild acidic pI. Each model protein showed a solubility minimum close to its pI, as expected, but the extent at which solubility was lost is dictated by NCPR. The higher the NCPR, the stronger the response in terms of loss of solubility. For instance, the highest propension to lose solubility at pI was observed for the highly acidic variant, which also presents the highest absolute value of NCPR. *Viceversa*,

the highly soluble protein at its pI was the mild acidic PNT variant and GFP, which have the lowest absolute value of NCPR. These data were confirmed by complementary biochemical and biophysical assays, indicating that the extent of solubility loss parallels that of aggregation increase. A similar behaviour observed for a globular and a disordered protein suggests that NCPR can have a general role in predicting the solubility properties of proteins. As regards IDP variants, the aggregation propensity observed for high-NCPR proteins behaviour was not predicted by dedicated bioinformatics algorithms, such as Zyaggregator and Aggrescan (Conchillo-Solé et al., 2007; Tartaglia and Vendruscolo, 2008). This is an indirect indication that the behaviour experimentally observed does not simply stems from the amino acid sequence of a protein, but from its interaction with an environment stimulus, such as pH that is not taken into account by the algorithms. Similar behaviours were observed when PNT variants were fused with GFP, which minimally contributes to the solubility of chimeras. These data suggest that the overall solubility of a protein can be driven by protein regions endowed with higher NCPR and, hence, prompter to respond to pH changes. This work has been already published (Tedeschi et al., 2017) and presented in the first section of following chapter entitled "Experimental work".

The issue of how charge patterning influences the compaction of IDPs is presented in the second section of "Experimental work" chapter. We explored here the effect of charge distribution on the conformational properties of two model proteins, N_{TAIL} and PNT4, endowed with similar NCPR (absolute value <0.05), FCR (~0.3), and hydrophobicity, but different proline content (~11% and ~5%, respectively). Each protein was rationally designed to obtain two permutants in order to sample different the lowest and the highest values of the parameter κ compatible with the natural amino acid composition. Then, the conformational properties of *wt* and κ-variants have been assessed through biochemical and biophysical techniques. As expected, experimental data show a direct correlation

between κ value and protein compaction. The protein variants which show more compact conformation are those featuring the highest κ values. Besides the increase in compactness, high-κ variants also show an increase in secondary structure content, which was not revealed by computational algorithms devoted to disorder prediction through sequence analyses (Pondr-fit meta-predictor). Moreover, the extent of response to charge clustering mainly reflects the content of proline residues. The higher the proline content, the lower the response in terms of compaction. The abundance of this amino acid emerges as a main cause of resilience to conformational compaction governed by charge patterns.

Both studies presented in this thesis take advantage from high *designability* of IDPs (Dunker et al., 2005b). This consists in the possibility to generate an ideally infinite series of *ad-hoc* sequences sharing some properties (*e.g.*, hydrophobicity, length, "depth" of structural disorder, etc) and differing in others (*e.g.*, net charge, charge density and distribution etc).

Both studies experimentally contribute to demonstrate that IDPs respond to the following rules: *i*) they lose solubility at their pI, as globular proteins do; *ii*) IDPs compactness is dictated by charge patterning. Actually, both studies help to define better the limits at which those rules are effective For instance, the solubility of low-NCPR proteins remains almost unaffected at pI. On the other side, the compaction effects of charge clustering risk to be almost undetectable in proline-rich IDPs. Overall, these results help to understand the sequence determinants of aggregation and conformational properties of IDPs. These data would greatly help in the *de-novo* design of synthetic, disordered solubility/aggregation tags and also to stimulate future studies aimed at rationally conceiving synthetic IDPs with a desired degree of solubility and compactness. To note that some of the properties experimentally unveiled for *re-designed* IDPs, such as the NCPR-driven propensity to aggregation, or the κ-related increase of compactness, were not anticipated by bioinformatics algorithms based on sequence analyses. This

suggests that the design of solubility/compactness properties requires the use of more complex informatics tools, such as those predicting the properties of conformational ensembles.

4. Experimental work

4.1 Aggregation properties of a disordered protein are tunable by pH and depend on its net charge per residue

Aggregation properties of a disordered protein are tunable by pH and depend on its net charge per residue

Giulia Tedeschi, Marco Mangiagalli, Sara Chmielewska, Marina Lotti, Antonino Natalello*, Stefania Brocca*

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

** Corresponding authors*

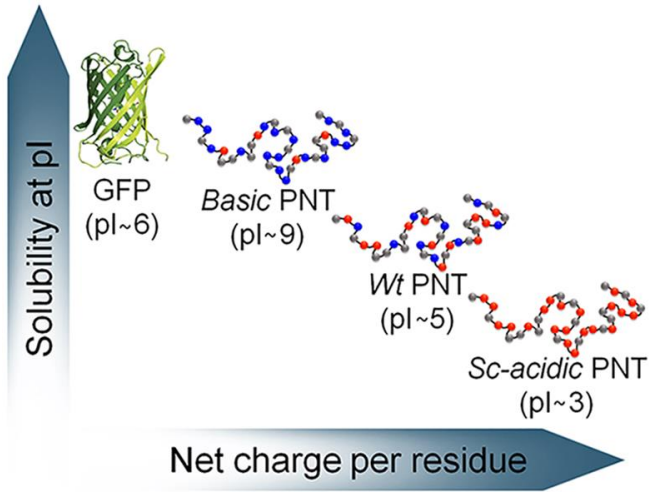
Department of Biotechnology and Biosciences, State University of Milano-Bicocca,

Piazza della Scienza 2, 20126, Milano, Italy

Keywords: IDPs, isoelectric point, NCPR, protein solubility, protein aggregation.

Abbreviations

ATR: attenuated total reflection; **CD**: Circular dichroism (spectroscopy); **FTIR**: Fourier transform infrared (spectroscopy); **GFP**: green fluorescent protein; **IMAC**: immobilized-metal affinity chromatography; **IDPs**: intrinsically disordered proteins; **FCR**: fraction of charged residue; **NCPR**: net charge per residue; **PB**: phosphate buffer; **PNT**: N-*terminus* moiety of measles virus phosphoprotein; **pI**: isoelectric point; **RC**: random coil.



Highlights

- Intrinsically disordered proteins lose solubility at isoelectric point (pI)
- The extent of solubility loss depends on net charge per residue (NCPR)
- Chimeric proteins with high- and low-NCPR moieties lose solubility at their average pI
- In chimeric proteins, high-NCPR moiety drives the loss of solubility

Abstract

Intrinsically disordered proteins (IDPs) possess a peculiar amino acid composition that makes them very soluble. Nevertheless, they can encounter aggregation in physiological and pathological contexts. In this work, we addressed the issue of how electrostatic charges can influence aggregation propensity by using the N-*terminus* moiety of the measles virus phosphoprotein, PNT, as a model IDP. Taking advantage of the high sequence *designability* of IDPs, we have produced an array of PNT variants sharing the same hydrophobicity, but differing in net charges per residue and isoelectric points (pI). The solubility and conformational properties of these proteins were analysed through biochemical and biophysical techniques in a wide range of pH values and compared with those of the green fluorescence protein (GFP), a globular protein with lower net charge per residue, but similar hydrophobicity. Tested proteins showed a solubility minimum close to their pI, as expected, but the pH-dependent decrease of solubility was not uniform and driven by the net charge per residue of each variant. A parallel behaviour was observed also in fusion proteins between PNT variants and GFP, which minimally contributes to the solubility of chimeras. Our data suggest that the overall solubility of a protein can be dictated by protein regions endowed with higher NCPR and, hence, prompter to respond to pH changes. This finding could be exploited for biotechnical purposes, such as the design of solubility/aggregation tags, and in studies aimed to clarify the pathological and physiological behaviour of IDPs.

1. Introduction

Protein aggregation is involved in a number of physiological and pathological events. Moreover, it is a major hurdle in the production and storage of recombinant proteins, included drugs. Hence, understanding the physical and chemical bases of protein aggregation could help not only to figure out how physio-pathological processes occur, but also to exploit this phenomenon for biotechnical purposes, for instance to increase *in-vitro* solubility of proteins (Trevino et al., 2008), to design biomaterials with tunable aggregation properties (Simon et al., 2017a), or even to design tags exploitable in the production of recombinant proteins (Costa et al., 2014).

How to recognize or predict protein solubility? Different definitions and criteria have been proposed, based on experimental observations, databases of soluble and insoluble proteins, or on the employment of machine-learning algorithms (Agostini et al., 2012; Niwa et al., 2009; Weiss et al., 2009). It emerges that besides sequence and structural features, the electrostatic properties of proteins, *i.e.* their net charge, can play a key role. It is well recognized that proteins behave as amphoteric molecules, showing significantly reduced solubility and even precipitation at their isoelectric points (pIs) (Loeb, 1918). On the other hand, charges can produce opposite effects. Indeed, "supercharging" of proteins, especially with negative charges, may enhance solubility (Su et al., 2007; Zhang and Liu, 2004), whereas positively-charged surface patches correlate with insolubility of proteins expressed in a cell-free *Escherichia coli* system (Chan et al., 2013). Systematic studies on protein solubility find obvious limitations in the disastrous structural effects induced by extensive replacement of charged residues on globular proteins. In this context, intrinsically disordered proteins (IDPs) provide a very versatile tool to extend the "host-guest approach" (Shoemaker et al., 1987) from peptides to larger molecules, minimizing structural effects. IDPs are usually well soluble proteins lacking strict spatial constraints and

compositional complexity (Dunker et al., 2001b; Li et al., 1996; Uversky, 2013c; Williams et al., 2001). Due to their high *designability* (Dunker et al., 2005a), starting from a "prototypical" sequence, it is possible to generate an ideally infinite series of *ad-hoc* proteins sharing some properties (*e.g.*, hydrophobicity, length, "depth" of structural disorder, etc) and differing in others (*e.g.*, net charge, charge density and distribution etc). IDPs have proved to be less prone to β-aggregation (Linding et al., 2004) and more stable to heat and pH than their folded counterparts (Campos et al., 2011; Csizmók et al., 2006). These features arise from the peculiar amino acid composition of IDPs and are consistent with the abundance of highly soluble residues (proline, charged and polar residues) and the paucity of aromatic and hydrophobic residues (Lise and Jones, 2005; Uversky et al., 2000; Weathers et al., 2004). These properties have suggested the use of IDPs as solubility enhancers, and the hypothesis they can act as "entropic bristles" sweeping the space around the fusion protein and preventing large molecules to participate in aggregation (Santner et al., 2012). Nevertheless, the high solubility of IDPs does not imply a lower propensity to collective interactions, such those giving rise to aggregates or coacervates. Indeed, besides hydrophobicity, also entropic factors, hydrogen bonding and electrostatic interactions can cause aggregation (Linding et al., 2004). Moreover, the water solubility of IDPs might depend on conformational compactness that, in its turn, is influenced by the water exposure of solubility-promoting amino acids (Van Der Lee et al., 2014). An attempt to rationalize the relationships between electrostatic charges and conformation of IDPs is represented by charge-hydropathy plots (Uversky et al., 2000) and, more recently, by diagrams of states. Through these latter empirical diagrams, conformational states of IDPs have been related to fraction of charged residue (FCR) and net charge per residue (NCPR) (Das and Pappu, 2013; Das et al., 2015; Mao et al., 2010).

In this complex scenario, we aimed to shed light on the role of charged amino acids on IDPs solubility at pI, using as a model the N-*terminus* moiety of measles virus phosphoprotein (PNT) (Karlin et al., 2002). We compared at various pHs the aggregation propensity of wild-type PNT (*wt* PNT) and synthetic variants of PNT with higher net charge and markedly more acidic or basic pI. We included in our study the green fluorescent protein (GFP), a globular protein very similar to *wt* PNT in terms of net charge and pI, but differing in NCPR, which is the worthiest parameter to compare the net charge of proteins of different length. Furthermore, we explored the ability of all PNT variants and GFP to reciprocally influence their solubility in chimeric constructs.

Our study shows that overall PNTs are more pH-responsive than GFP, which has lower NCPR. Among PNT variants, the loss of solubility occurs to varying degree, depending on the protein net charge. PNT variants endowed with highest NCPR promptly undergo aggregation at or near their pI, whereas low-NCPR proteins mildly react to pH, remaining mostly soluble. We further report that PNT variants "transmit" their solubility profile to chimeric constructs with GFP. This information would greatly help in the *de-novo* design of synthetic, disordered solubility/aggregation tags and hopefully in understanding *in-vivo* processes of IDP condensation and aggregation.

2. Materials and methods

2.1 Gene design and cloning

Wild-type PNT (*wt* PNT) was cloned in pET-21a [PNT] vector (Sambi et al., 2010). Acidic and basic variants of PNT were obtained through gene synthesis (Genscript, Piscataway, NJ, USA). Two kinds of *supercharged* (*sc*) variants of PNT were designed. In the *sc-acidic* PNT, His, Lys and Arg residues of *wt* PNT were substituted with either Glu or Asp; the *basic* variant and the *sc-basic* PNT

variants were obtained by substitution of Glu and Asp residues with Lys and Arg residues. Synthetic genes were cloned into pET-21a vector (EMD, Millipore, Billerica, MA, USA), between the sites *Nsi*I and *Not*I, giving rise to plasmids pET-21a [*sc-acidic* PNT], pET-21a [*basic* PNT], pET-21a [*sc-basic* PNT]. In this work, we indicate as pET-21a [PNTs] the ensemble of expression vectors carrying aforesaid PNT genes.

Constructs for the fusion of GFP at the C-*terminus* of PNT mutants were obtained by cloning the GFP gene into pET-21a [PNTs] digested with *Nde*I. The coding sequence was amplified by PCR from pET-19b [GFP](Sambi et al., 2010) with primers inserting *Nde*I restriction sites at both 5' and 3' extremities. The forward and reverse primers for amplification were: FW 5'- GGATCCCATATGAAAGTGAGCAAG - 3', RV 5'- CATATGCCCAAGCTTCTTGTACAG -3' (*Nde*I restriction site is underlined). Amplification reactions were carried out using Q5® High-Fidelity DNA Polymerase (New England Biolabs, Ipswich, MA). The reaction conditions used were: 1 cycle (98° C for 2 min), 25 cycles (98° C 10 sec, 56° C 30 sec, and 72° C 1 min), and a final cycle of 72° C 3 min. The PCR product was preliminarily cloned into pUC18 blunt-end digested with *Sma*I obtaining pUC18 [GFP]. The GFP gene was then excised from pUC18 [GFP] digested with *Nde*I and gel-purified before ligation into the pET-21a [PNTs] cleaved with the same restriction enzyme.

The correct orientation of the GFP insert in the pET-21a [PNTs-GFP] vectors was verified by enzyme restriction and by bidirectional DNA sequencing. The amino acid sequences of PNTs are reported in **Figure 1**. GFP was produced from pET-19b [GFP] (Sambi et al., 2010).

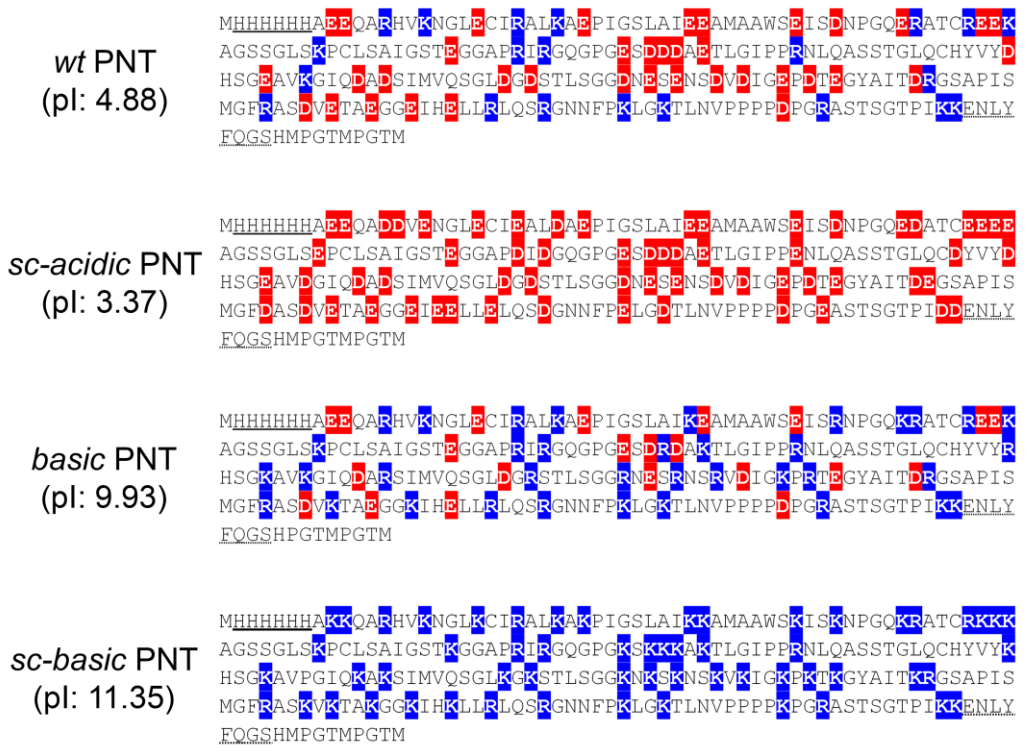


Figure 1. Sequence of PNTs variants. For each sequence, acidic and basic residues are highlighted in red and blue respectively. 6xHis tag is underlined in continuous line and TEV protease site is underlined in dotted line.

2.2 Protein production and purification

Escherichia coli strain BL21[DE3] (EMD, Millipore, Billerica, MA, USA) was used as the host for heterologous production of PNTs variants. Transformed cells were grown overnight at 37° C in Lennox medium (10 g/L tryptone, 5 g/L yeast extract, 5 g/L NaCl), diluted 1 : 20 in 200 mL of Zym-5052 medium (Studier, 2005) and incubated at 25° C. Media were added of 100 mg/L ampicillin.

Proteins were extracted as described in (Parravicini et al., 2015) and recombinant, his-tagged proteins were purified by immobilized-metal affinity chromatography (IMAC) on Ni/NTA agarose gel (Jena Bioscience, Jena, Germany) at 4° C. To

improve the purification yield, clarified lysates were incubated at 4° C for 1 hour with Ni/NTA agarose gel before purification.

Protein concentration was determined by the Bradford protein assay (Bio-Rad, California, USA), using bovine serum albumin as a standard.

Samples containing highest protein concentrations were buffer exchanged twice by gel filtration on PD10 column (GE Healthcare, Little Chalfont, UK) against 10 mM ammonium acetate buffer pH 7.0.

2.3 Biochemical and biophysical analyses

Since pH strongly impacts on protein solubility and affects determination of protein concentration by Bradford assay, samples were prepared by a procedure allowing to minimize differences in protein yield and sample concentration. After buffer exchange in 10 mM ammonium acetate, elution fractions containing protein at the highest concentrations were pooled, newly quantified and divided in samples containing the same protein amount. Samples were lyophilized in a freeze-dryer (Heto FD1.0 Gemini BV, Apeldoorn, Netherlands) and then suspended in equal volumes of 10 mM potassium phosphate buffer (PB) at different pH values (3.0, 5.0, 6.0, 7.0, 9.0). Only GFP, PNT *basic* and PNT *basic*-GFP were assayed also at pH 8.0, 8.5, 9.5, 10.0 and 11.0; while *sc-acidic* PNT were further assayed at pH 4.0. The pH measurements were carried out at room temperature with a HI 9321 Microprocessor pH meter (Hanna Instruments, Italy). The instrument was calibrated against the standard pH 4.00 and 7.00 solutions (Sigma Aldrich, St. Louis, MO, USA).

2.3.1 Far-UV circular dichroism (*CD) spectroscopy*

Lyophilized samples were suspended in PB (0.09 mg/ml for GFP and PNTs variants and 0.18 mg/ml for fusion proteins) at different pH values, and incubated for 1 hour at room temperature. CD spectra were recorded at room temperature by a spectropolarimeter J-815 (JASCO Corporation, Easton, USA) in a 1-mm path-length cuvette. Measurements were performed at variable wavelength (190–260 nm) with scanning velocity 20 nm/min, bandwidth 1 nm, digital integration time per data 2 sec and data pitch 0.2 nm. All spectra were averaged from two independent acquisitions, corrected for buffer contribution, and smoothed by Means-Movement algorithm. Experiments were performed in triplicate.

2.3.2 Fourier transform infrared (FTIR) spectroscopy

Lyophilized samples were suspended in PB (1.5 mg/ml) at different pH values and incubated for 1 hour at room temperature. Two µl of the above protein solutions were deposed on the single reflection diamond element of the attenuated total reflection (ATR) device (Quest, Specac, USA) and dried at room temperature to obtain a protein film (Parravicini et al., 2013) [33]. The protein film on the ATR element was hydrated by adding 6 µL of D_2O close to the sample (Goormaghtigh et al., 1999; Natalello et al., 2015) and incubated for 1 hour at room temperature. The ATR/FTIR spectra were collected at room temperature using a Varian 670-IR spectrometer (Varian Australia Pty Ltd, Mulgrave, Victoria, Australia), equipped with a nitrogen-cooled Mercury Cadmium Telluride detector, under the following conditions: resolution 2 cm^{-1}, scan speed 25 kHz, 1000 scan co-additions, triangular apodization, and dry-air purging.

ATR/FTIR absorption spectra were corrected for buffer contribution, normalized at the Amide I' (1700-1600 cm^{-1}) band area and were smoothed using the Savitsky-Golay method before second derivative calculation. Spectral analyses were performed with the Resolutions-Pro software (Varian Australia Pty Ltd,

Mulgrave, Victoria, Australia). At least three independent measurements were performed for each condition.

2.3.3 Solubility assay

SDS-PAGE was used to assess solubility of GFP and PNT variants after incubation at different pH. Lyophilized samples were suspended in PB at different pH values, at concentration of 0.5 mg/mL. After 1-hour incubation, an aliquot was collected (total protein), and the remaining was centrifuged for 10 min at 15,000 x *g* to separate soluble and insoluble protein fractions. An equal volume (20 µl) of total and soluble proteins were separated in 14% SDS-PAGE (Laemmli, 1970) and stained with Gel-Code Blue (Pierce, Rockford, USA). Broad-range, pre-stained molecular-weight markers (GeneSpin, Milan, Italy) were used as standards. Densitometric volume of each protein band was calculated by the software Image Lab (Bio-Rad, California, USA). For each pH value, the relative amount of soluble protein (solubility) was calculated with reference to total protein in the aliquot. Percentages are referred to the highest value of solubility considered as 100%. Data represent an average of three independent biological replicates. Similar results were obtained from solubility tests carried out after 30 min-, 1 h- and 2 h-incubation for PNT variants and for GFP, at pH near the pI of each protein (data not shown).

2.4 Bioinformatic analysis

The theoretical pI was calculated with different algorithms: Expasy ProtParam (http://web.expasy.org/protparam) and Isoelectric point calculator (Kozlowski, 2016). Disorder prediction with Pondr-fit (Xue et al., 2010a) and plots of mean net charge versus mean hydropathy (Uversky et al., 2000) were used to assess conformation profile.

NCPR values were calculated as:

$$NCPR = \frac{(aa\ positive - aa\ negative)}{aa\ total}$$

FCR values were calculated as:

$$FCR = \frac{(aa\ positive) + (aa\ negative)}{aa\ total}$$

where *aa positive* is the number of positively charged amino acids, *aa negative* is the number of negative charged amino acids and *aa total* is the total number of amino acids (Mao et al., 2010).

NCPR, FCR and the Kyte-Doolittle hydropathy score (scaled from 0, least hydrophobic, to 9, most hydrophobic) were calculated through the webserver CIDER (Holehouse et al., 2017).

3 Results and discussion

Studies on aggregation/solubility of proteins are very challenging if we consider the faceted role different amino acid residues can have, depending on their physicochemical classes, solvent exposition, and on their position in a protein structure (Isom et al., 2011). Although IDPs have to be considered as conformational ensembles, their use as a model allows to greatly simplify the issue, as it allows to reduce the relevance of conformational effects and to focus on the "chemical behaviour" of "biological objects". Moreover, the relaxed conformational constrains on IDPs primary structure made it possible to design a "family" of related proteins that can be assimilated to ionisable amphoteric polyelectrolytes, whose response to chemical and physical laws can be gathered more easily than from a single protein.

3.1 Design of PNT variants and of their fusions with the green fluorescent protein (GFP)

To study systematically the solubility of a disordered protein, PNT variants were designed exploring a wide range of pI values and net charges. More in detail, our experimental approach was aimed at sampling two mild-charged and two *supercharged* (*sc*) basic and acidic variants of PNT. Since *wt* PNT already exhibits mild-acidic features, we designed three synthetic variants, thereof one is mild basic (simply referred as *basic*), one *sc-acidic* and one *sc-basic*. In the following, the ensemble of PNT variants used in this work is referred to as "PNTs". Overall, the design of synthetic PNT variants was carried out by reversing the sign of charged residues already present in the wild-type sequence while keeping unchanged all other residues (**Figure 2A)**. For this reason, all PNTs have a very similar fraction of charged residues (FCR, 0.257 ± 0.004) and hydropathy score (3.826 ± 0,067), as calculated by CIDER (Holehouse et al., 2017), and as shown in the Uversky plot (Uversky et al., 2000) (**Figure 2B**).

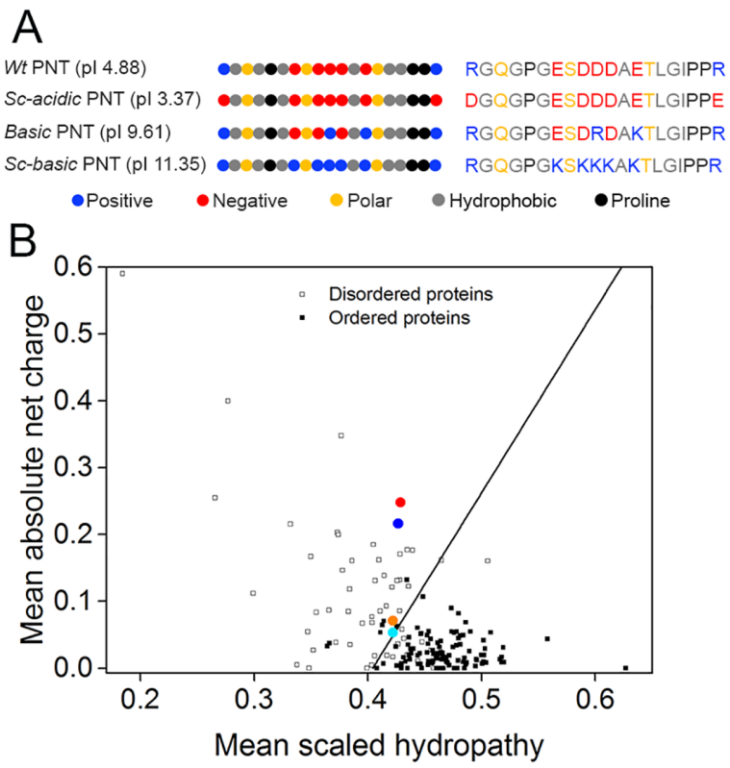


Figure 2. Design and disorder prediction of PNT variants. (A) Scheme of amino acid composition of PNT variants. Residues 85–105 are shown to exemplify the results obtained substituting charged residues in the same relative positions and keeping unchanged polar, hydrophobic and proline residues. **B)** Predictions were carried out using charge-hydropathy (Uversky et al., 2000)

Wild-type PNT was described previously (Sambi et al., 2010). Briefly, it spans the first 230 amino acid residues of the whole P protein and carries an N-terminal 6xHis tag and a C-terminal tail containing the TEV protease target sequence. This sequence contains 21 positively charged amino acid residues and 38 negatively charged residues resulting in an acidic pI of 4.88 and featuring an NCPR of - 0.071. The mild-charged *basic* variant was obtained by almost inverting the ratio of positively (37) and negatively (23) charged amino acid residues of *wt* PNT, and reaching a pI of 9.61 and NCPR of + 0.055. The *sc-acidic* PNT has a pI of 3.37

and includes 62 negatively charged residues (0 positives ones), with an NCPR of - 0.248, while *sc-basic* PNT has a pI of 11.44 and includes 57 positively charged residues (0 negatives ones), with an NCPR of + 0.216. We assumed that the 6xHis tag and TEV site affect all variants in the same way, producing effects negligible in the comparative analyses. The features of PNTs are summarized in **Table 1**, amino acid sequences are reported in **Figure 1** and plots of linear NCPR in **Figure 3**.

Protein ID	Amino acid content					pI	NCPR
	Lys	**Arg**	**His**	**Glu**	**Asp**		
Wt PNT	9	12	11	23	16	4.88 ± 0.03	- 0.071
Sc-acidic PNT	-	-	8	34	29	3.37 ± 0.06	- 0.248
Basic PNT	16	21	11	15	8	9.61 ± 0.37	+ 0.055
Sc-Basic PNT	45	12	11	-	-	11.35± 0.15	+ 0.216
Wt PNT-GFP	31	18	21	39	35	5.24 ± 0.06	- 0.048
Sc-Acidic PNT-GFP	22	6	18	50	48	4.16 ± 0.05	- 0.139
Basic PNT-GFP	38	27	21	31	27	8.45 ± 0.40	+ 0.014
Sc-Basic PNT-GFP	67	18	21	16	19	10.20 ± 0.25	+ 0.099
GFP	22	6	10	16	19	6.15 ± 0.09	- 0.023

Table 1. Features of proteins assayed in this work. The amino acid sequences are reported in **S1** and include 6xHis tag and TEV site. Along this paper, we will simply refer to the mean value of pI.

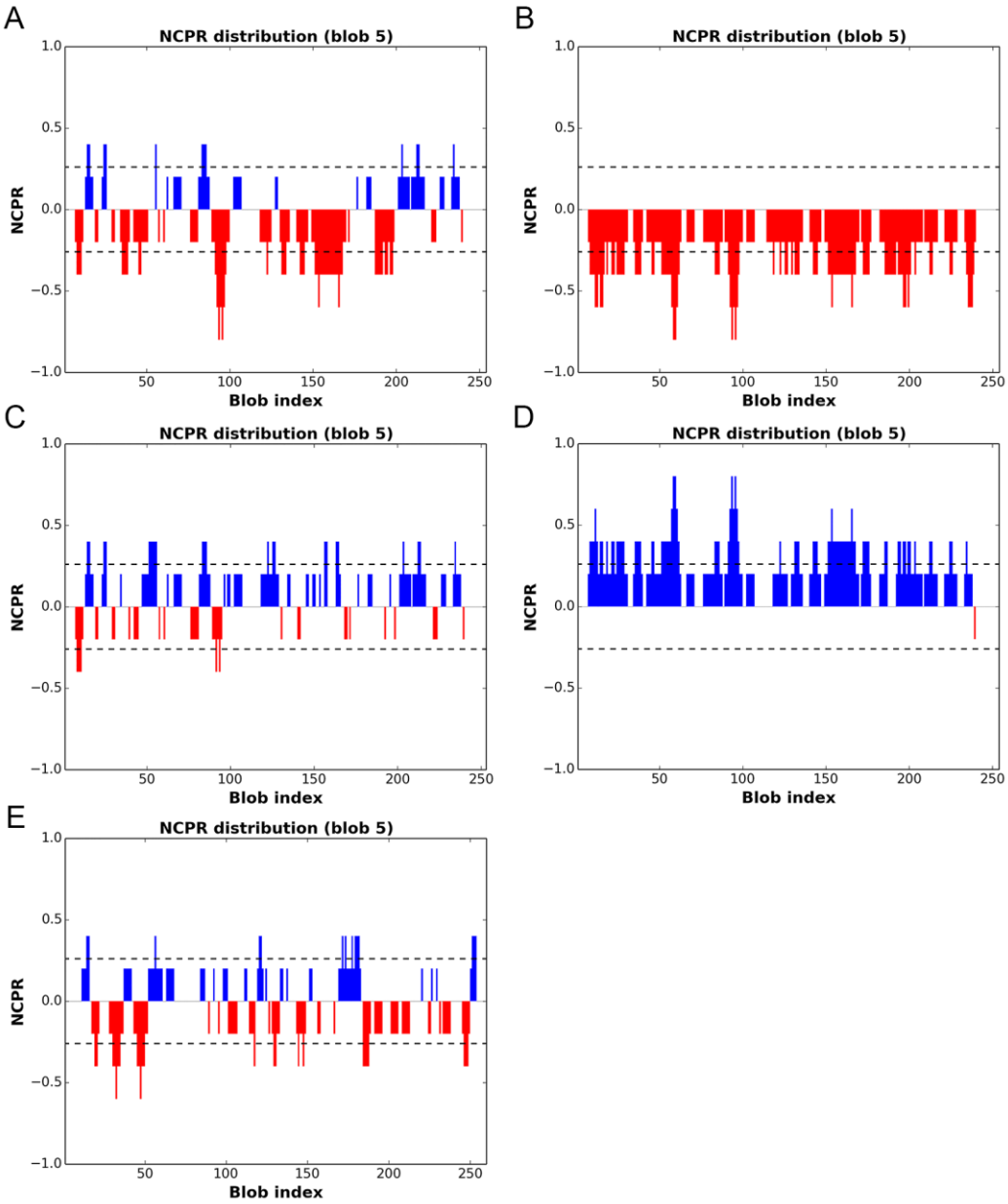


Figure 3. Linear NCPR plots. Calculations were carried out using CIDER webserver [37], with a sliding-window ("blob") of five residues. Blue and red denote positive and negatively charged residues, respectively. **A**) *wt*, **B**) *sc-acidic*, **C)** *basic*, **D)** *sc-basic* variants of PNT; **E)** GFP.

Despite the profound sequence changes so far described, the overall disorder profile of synthetic PNTs calculated by Pondr-fit (Xue et al., 2010a) remains similar to that of *wt* PNT, and slightly more disordered for the two *sc*-PNTs (**Figure 4**).

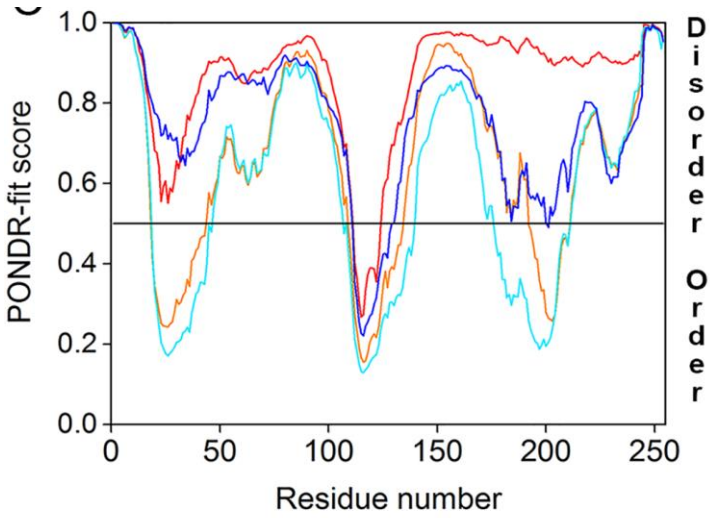


Figure 4. Disorder prediction of PNT variants. *Wt* PNT, *sc-acidic* PNT, *basic* PNT and *sc-basic* PNT are indicated in orange, red, light blue and blue, respectively.

Each PNT variant was C-terminally fused to GFP to assay the ability of each moiety to affect the solubility of the fusion partner. The GFP shares with PNTs a very similar hydropathy score (3.94) and FCR (0.246), but has lower NCPR (- 0.023). Features of PNTs fused with GFP are included in **Table 1**. All proteins but *sc-basic* PNT and *sc-basic* PNT-GFP were produced in Zym 5052 medium and purified at comparable yield (~ 4 mg per liter of colture) from the soluble fraction of cell extracts. In the case of *sc-basic* PNT and *sc-basic* PNT-GFP, we did not observe any production of the recombinant proteins, even in the insoluble protein fraction. This problem has been already referred by other Authors for a supercharged globular protein (Lawrence et al., 2007). We can hypothesize that

the high frequency of Lys and Arg residues in *sc-basic* PNT sequence may unfavourably impact on its translation rate, hence producing ribosome stalling and transcript degradation (Charneski and Hurst, 2013). Further attempts to produce other *sc-basic* PNTs with slightly modified sequences were unsuccessful. Since *sc-acidic* PNT, *wt* PNT and *basic* PNT allow us to sample a wide range of NCPR and pI values, we considered this ensemble of proteins, along with GFP, suitable to test our hypothesis. All proteins were purified by IMAC, liophilyzed, and resuspendend in phosphate buffer (PB) adjusted at different pH and finally solubility was assessed. All samples analysed before and after lyophilisation gave superimposable spectra of far-UV CD and FTIR spectroscopies (data not shown).

3.2 Solubility and propensity to aggregation of PNT variants and GFP

The solubility at different pHs of PNT variants and GFP was studied *in vitro* using three complementary techniques: solubility assays, far-UV CD and FTIR spectroscopies. We have considered a "standard range" of pH values (3.0, 5.0, 6.0, 7.0, 9.0) to analyse all the proteins and compare at a glance their solubility profiles. Other pH values were chosen *ad hoc* to study more extensively some of the proteins (see later).

The CD spectrum of *wt* PNT at pH 7.0 is that typical of a disordered protein, with a deep downward peak in the range 190-200 nm **(Figure 5A)**. The shape of this spectrum is consistent with that already published for the same protein and measured in sodium phosphate buffer at pH 7.5 (Sambi et al., 2010). The ellipticity value observed at 222 nm is consistent with the existence of some residual helical structure. Overall, at pH 7.0, PNTs spectra are similar as for profile and ellipticity. As the pH reaches the pI value of each PNT variant, we observed a dramatic loss of the ellipticity signals (**Figure 5 A-C**). Moreover, in the *sc-acidic* PNT sample we detected an increase of ellipticity at 190 nm and a shift of the minimum toward 218 nm, suggesting the simultaneous formation of β-structure.

Overall, far-UV CD spectroscopy analyses hint that PNTs undergo aggregation as pH approaches to their pI. Solubility assay and FTIR analyses were performed to assess this hypothesis.

Solubility was quantified by densitometric analysis of samples after SDS-PAGE separation. We detected the lowest solubility of *wt*, *basic* and *sc-acidic* PNT at pH 5.0, 3.0 and 9.0, respectively (**Figure 5 D-F**). This observation is in good agreement with the flattening of CD signal observed under the same conditions. It is worth to notice that the decrease in measured solubility is higher for *sc-acidic* PNT (~ -95%) than for *wt* (~ -60%) and *basic* PNT (~ -50%).

The FTIR second derivative spectra (whose minima correspond to absorption maxima) of PNTs were reported in the Amide I' band in **Figure 5 G-I**. We show here spectra obtained after H/D exchange, since they allow to better resolve the spectral signature of different structural secondary elements and to distinguish between α-helical and disordered structures.

A scheme of the typical absorption regions of the different protein secondary structures for samples in D_2O is explicitly reported in the spectrum of *wt* PNT in **Figure 5 G** (Barth, 2007; Natalello et al., 2012). The FTIR second derivative spectra of PNTs at pH 7.0 show a main component around 1641 cm^{-1} **(Figure 5 G-I)** that can be assigned to disordered structures.

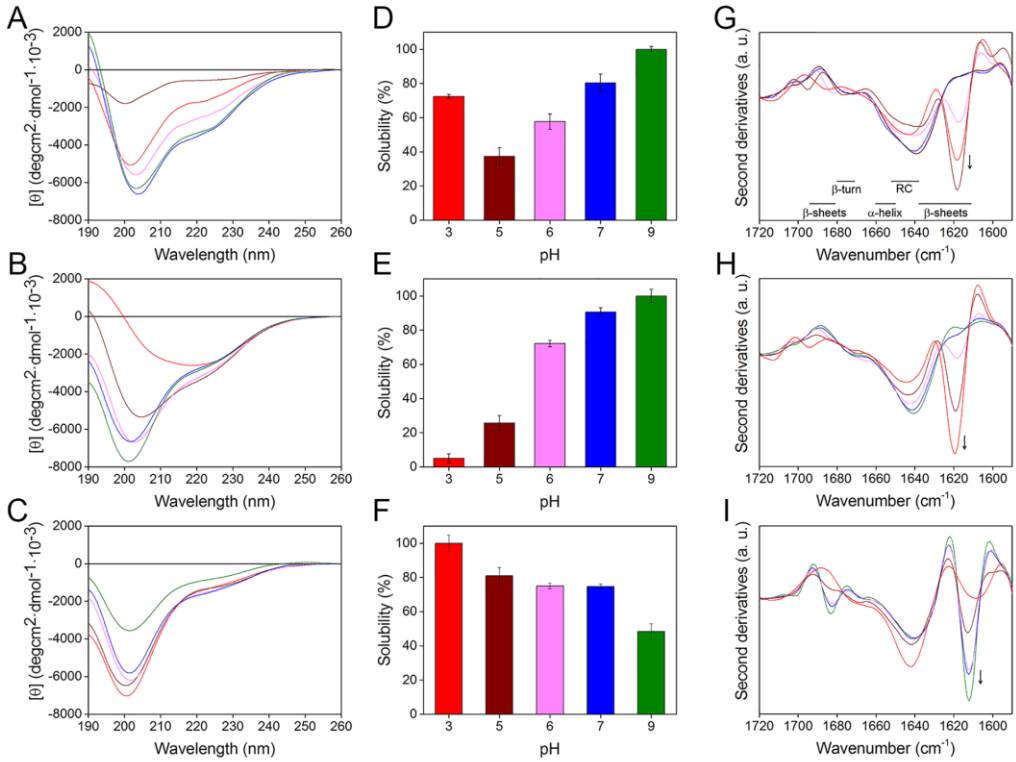


Figure 5. Solubility and propensity to aggregation of single PNTs. For each analysis, proteins were prepared in PB at pH 3.0 (red), 5.0 (brown), 6.0 (pink), 7.0 (blue) and 9.0 (green). Upper, middle and lower row refers to *wt*, *sc-acidic* and *basic* PNT, respectively. **A-C)** Far-UV CD spectra. It is shown one of three independent experiments. **D-F)** Solubility assay. Error bars indicate standard deviations on three independent experiments. **G-I)** Second derivatives of the FTIR absorption spectra. Arrows point to increasing intensity of the intermolecular β-sheet peak. The Amide I' band assignment to the protein secondary structures is also given in **G.** RC: random coil. It is shown one of three independent experiments.

According to solubility assays, spectra collected at different pHs show an additional component around 1619-1613 cm^{-1} (arrow in Figure 2.G-I), whose intensity increases as pH reaches the pIs of the different PNTs and indicates the formation of intermolecular β-sheets (Barth, 2007; Natalello et al., 2012).

For the sake of completeness, we also measured the solubility of GFP at different pHs. GFP is a globular protein endowed with a well-defined β-barrel structure composed of 11 β-strands (Yang, 1997), and a theoretical pI of 6.15. CD spectra between pH 6.0 and 9.0 show a positive peak at 195 nm and a broad negative peak at 218 nm, as expected for a natively structured protein with a predominant content of β-strands (**Figure 6 A**). At pH 5.0, GFP reaches its lowest solubility, with a moderate loss of the CD signal and comparable loss of soluble protein (~ -20%) (**Figure 6 A, B**). The difference between the observed pH dependence and the theoretical pI of GFP may be due to pKa shifts of titratable residues, which, in turn, may depend on their positions in the *core* of a folded protein (Isom et al., 2011). At pH 3.0, GFP is partially unstructured (**Figure 6 A**) and yet soluble (**Figure 6 B**). The structural transitions of GFP were confirmed and completed by FTIR analyses (**Figure 6 C**). The second derivatives of the IR absorption spectra at pHs 6.0-9.0 show a main component at ~1623 cm^{-1} that, along with the peak around 1689 cm^{-1}, is due to native intramolecular β-sheets. At lower pHs a partial loss of the native components indicates protein unfolding, which is more evident at pH 3.0 (**Figure 6 C**).

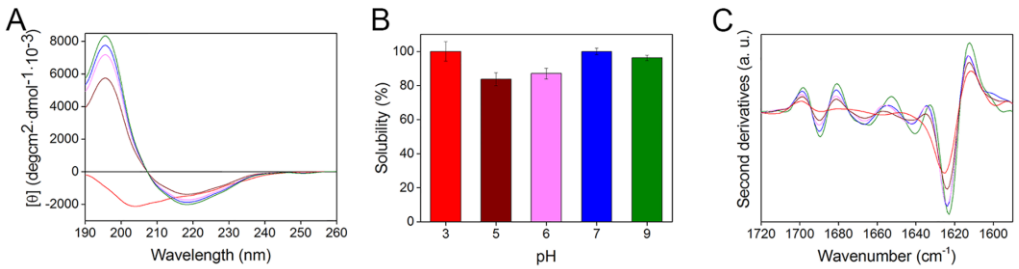


Figure 6. Solubility and propensity to aggregation of GFP. For each analysis, GFP was in PB at pH 3.0 (red), 5.0 (brown), 6.0 (pink), 7.0 (blue) and 9.0 (green). **A)** Far-UV CD spectra. It is shown one of three independent experiments. **B)** Solubility assay. Error bars indicate standard deviations on three independent experiments. **C)** Second derivatives of the FTIR absorption spectra. It is shown one of three independent experiments.

Taken together, these results highlight that changes of pHs produce a stronger impact on the solubility of PNT variants than on the solubility of the globular GFP. What makes the difference in the behaviour of GFP and PNTs? It is well reasonable to assume that compaction and folding may influence protein solubility through the exposure at different extent of solubility-promoting residues (Kramer et al., 2012). However, we considered that other protein features might be of relevance, in particular the difference in the net charge per residue that is described by NCPR (Mao et al., 2010). When challenged at different pHs, our set of proteins aggregate at or near their pI, with different intensities which reflect the absolute value of NCPR ($NCPR^{sc\ acidicPNT} = \left| - 0.248 \right| > NCPR^{wtPNT} = \left| - 0.071 \right| > NCPR^{basicPNT} = \left| + 0.055 \right| > NCPR^{GFP} = \left| - 0.023 \right|$). Among PNTs, we observed the strongest pH-dependent aggregation with *sc-acidic* PNT, whereas the loss of solubility of *basic* PNT was the mildest. According with the reported results, we concluded that NCPR should be taken into careful consideration to predict pH-dependent aggregation. This interpretation is indirectly corroborated by experimental data on the high and pI-

independent solubility in the pH range 2-12 of charge-free proteins, which obviously exhibit null NCPR (Højgaard et al., 2016).

3.3 Solubility and aggregation of GFP fused to PNT variants

To investigate the behaviour of PNTs as solubility tags, we performed the same experiments described above with chimeric proteins composed by PNT variants and GFP (**Table 1)**. The CD spectra at pH 7.0 of all chimeras (**Figure 7 A-C**) are similar to those already observed for *wt* PNT-GFP in similar conditions (PB at pH 7.5) (Sambi et al., 2010), with a negative peak at 205 nm, instead of the deep downward peak typically observed around 190-200 nm in disordered proteins. When pH approaches the pI of the respective IDP moieties, a marked spectral flattening occurs. This observation is consistent with solubility profiles, which roughly parallel those observed for respective individual PNTs in the same pH range (**Figure 7 D-F**).

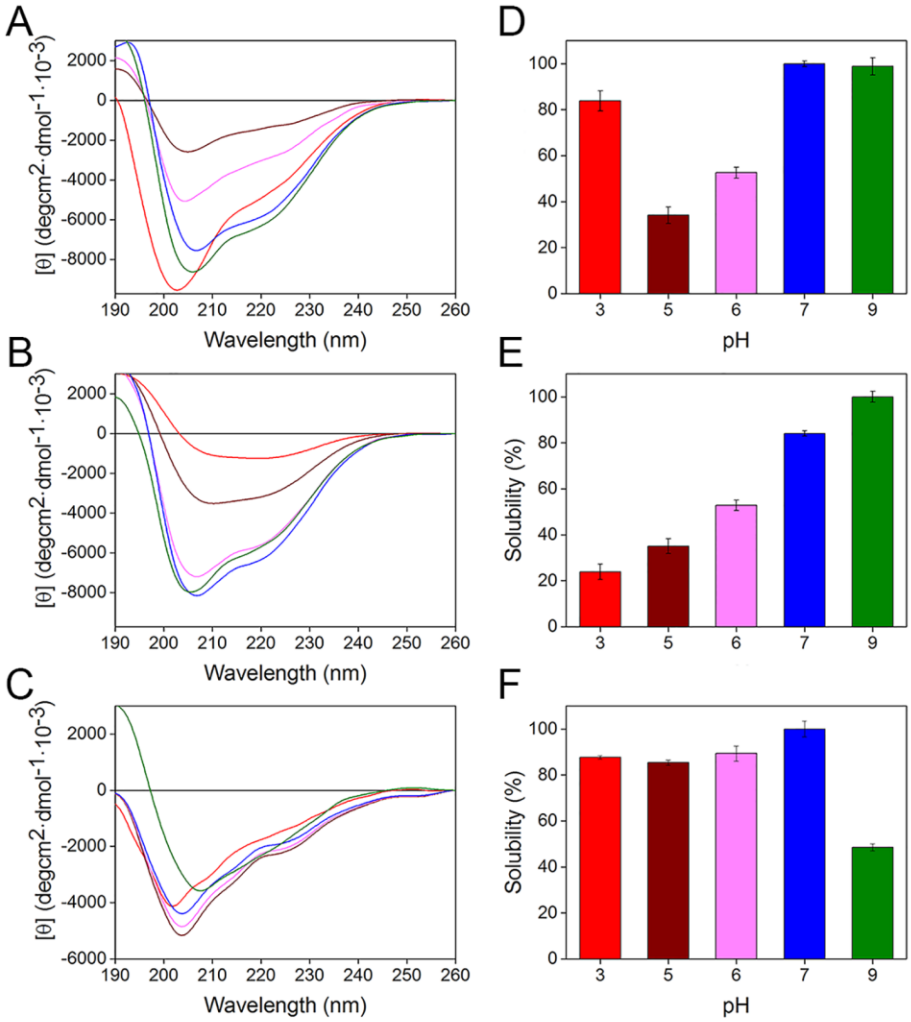


Figure 7. Solubility and aggregation of GFP fused to PNT variants. For each analysis, proteins were in PB at pH 3.0 (red), 5.0 (brown), 6.0 (pink), 7.0 (blue) and 9.0 (green). Upper, middle and lower row refers to *wt, sc-acidic and basic* PNT, respectively. **A-C)** Far-UV CD spectra. It is shown one of three independent experiments. **D-F)** Solubility assay. Error bars indicate standard deviations on three independent experiments.

It is worth to remark that *sc-acidic* PNT-GFP undergoes the most intense loss of solubility (~ -95%) at pH 4.0 (data not shown), likely reflecting the pI of the chimeric protein (4.16), rather than the pI of the lone PNT moiety (3.37). Such a pI shift is hard to be experimentally detected in *wt* PNT and its GFP fusion because of the proximity of their pIs (4.88 and 5.24, respectively). The behaviours of *basic* PNT and its GFP-fusion were similar even in the range of pH 8.0 - 11.0 (**Figure 8**).

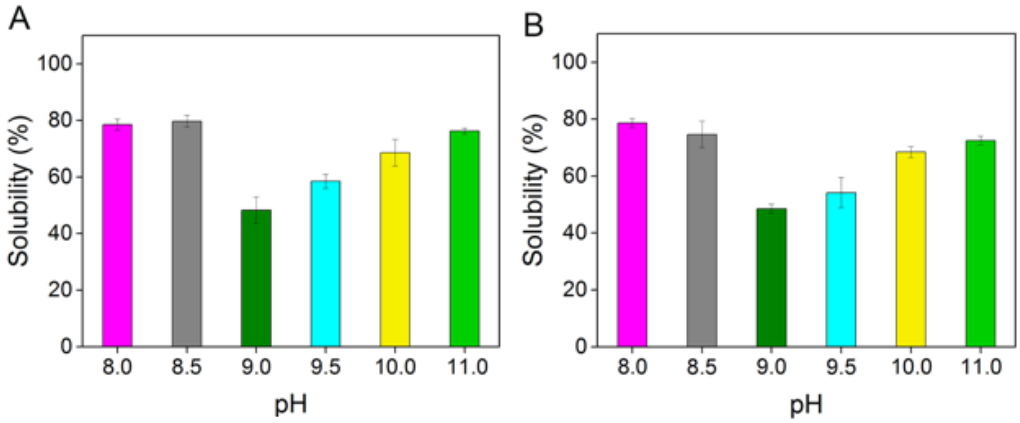


Figure 8. Solubility assay of *basic* PNT *per se* (A) and fused to GFP (B). For each analysis, proteins were dissolved in PB at different pHs. Error bars indicate standard deviations on three independent experiments.

Although it is difficult to generalize, we can consider that small pI differences are hardly detectable in a shallow solubility profile, as that of *basic* PNT, *vice versa* they are strikingly evident in systems that are more pH-sensitive, as that of *sc-acidic* PNT.

The FTIR second derivative spectra of the GFP fusions at pH 7.0 (**Figure 9 A-C**) mainly show the sum of spectral components observed for the isolated GFP and PNT variants at the same pH. FTIR spectra are consistent with the data on solubility, since the intensity of the intermolecular β-sheet component (~1619-

1613 cm^{-1}) increases as the pH approaches the pI of the disordered moiety (**Figure 9 A-C**).

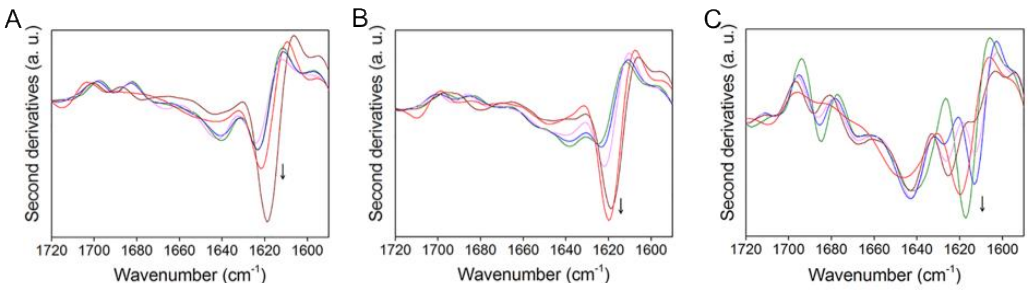


Figure 8. Solubility assay of *basic* PNT *per se* (A) and fused to GFP (B). For each analysis, proteins were dissolved in PB at different pHs. Error bars indicate standard deviations on three independent experiments.

Since solubility/aggregation profiles of single PNTs and their GFP-fused counterparts are very similar, one can reason that GFP exerts a marginal effect on the overall solubility of chimeric proteins.

From our results, we can infer that PNT variants with the highest NCPR, *i.e. wt* and *sc-acidic* variants, are able to prime the aggregation burst of whole chimeric constructs, suggesting the use of similar IDPs as aggregation tags rather than as solubility tags. *Vice versa*, milder charged polyampholytes, i.e. *basic* PNT, are less sensitive to pH changes and can cope with a broad range of pHs without undergoing aggregation.

This information may help to better define and to rationalize the properties of an effective solubility enhancer, already described in the pivotal work of Santner et al. 2012 as an entropic bristle of similar size and different pI than the target protein (Santner et al., 2012). Overall, our results indicate that the use of supercharged proteins as solubility enhancers is inherently risky, since high net charge, besides driving extremely high solubility, can also lead to extensive aggregation.

Moreover, data reported suggest that each moiety of a fusion protein may "sense" environmental pH according to its own features. When NCPR is uneven along a sequence, "local" values of NCPR should be considered instead of a whole NCPR score, averaged on the entire sequence. For instance, the low NCPR value of *basic* PNT-GFP (+ 0.014) would induce to underestimate its propensity to aggregate driven by the disordered partner (NCPR= + 0.055). Plots of linear NCPR can be useful to see at a glance the distribution of charged residues along a sequence (see **Figure 3**).

Can we generalize the behaviour observed for GFP and PNT to other globular and disordered proteins? We can reason that pH-sensitivity can be exacerbated in protein regions where one type of charged residue is recurrent (*e.g.* arginine-rich protamines; Glu/Asp-rich prothymosin α). Such low-complexity sequences are thought to enable IDPs to undergo fast, collective interactions (Brangwynne et al., 2015; Halfmann, 2016). Condensation of protein-rich assemblies has been recognized to foster liquid-liquid phase transitions giving rise to functionally important, membrane-less subcellular compartments, such as nucleoli, RNA granules, Cajal bodies. It is conceivable that different pH-sensitivity may impart different aggregation propensity and "phase behaviour", in response to even subtle changes of intracellular pH or NCPR. Indeed, transitions from expanded coil to collapsed globules often occur suddenly and can be reversed by even small changes in the NCPR (Das and Pappu, 2013; Das et al., 2015), suggesting the existence of a threshold value of this parameter which delimits the two conformational ensembles. To conclude, it seems that we can still learn a lot by reconsidering and applying long-time known chemical-physical principles to new questions, such as the aggregation/coacervation of disordered proteins. The high designability of IDPs will help to experimentally prove and further understand mechanisms that may in general influence the aggregation of proteins.

4. Conclusions

We found that aggregation propensity in a set of model proteins mainly responds to pH changes according to NCPR absolute value. Besides the expected loss of solubility at pI, we found that "aggregation intensity" is directly proportional to NCPR, which correlates net charge to protein size. This implies that proteins endowed with similar net charge and pI can behave differently in terms of "aggregation intensity", according to their NCPR. Moreover, protein regions with highest NCPR leads the overall behavior in chimeric proteins.

The overall rules dictating the aggregation appear captivating in their simplicity, in spite of the complexity of physiological and pathological phenomena in which might be involved. These observations may contribute to understand the behavior of IDPs in response to events (e.g., post-translational modifications, environment pH changes, mutations) that can affect protein NCPR. Moreover, this knowledge can have applicative potential in the design of solubility/aggregation tags for recombinant proteins.

Acknowledgments

This work was partly supported by a grant Fondo di Ateneo (FA) of the University of Milano-Bicocca to SB, AN and ML.

4.2 Clustering of charged residues and proline content affect conformational properties of intrinsically disordered proteins

Clustering of charged residues and proline content affect conformational properties of intrinsically disordered proteins

Giulia Tedeschi[1], Edoardo Salladini[2], Carlo Santambrogio[1], Rita Grandori[1], Sonia Longhi[2*], Stefania Brocca[1*]

[1]*Department of Biotechnology and Biosciences, State University of Milano-Bicocca, Piazza della Scienza 2, 20126, Milano, Italy*

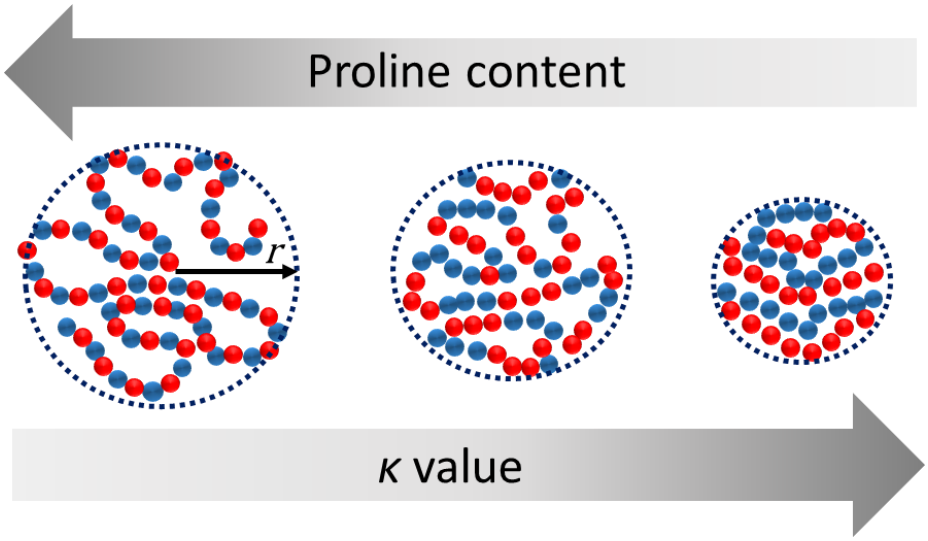
[2]*Aix-Marseille Univ, CNRS, Architecture et Fonction des Macromolécules Biologiques (AFMB), UMR 7257, 13288 Marseille, France*

**Corresponding Authors*

Keywords: IDPs, compaction index, κ value-normalized compaction index, small-angle X-ray scattering, net charge per residues (NCPR), charge distribution, κ value.

Abbreviations:

CD: circular dichroism (spectroscopy); **CI**: compaction index; **$^{\kappa}$CR**: κ-normalized compaction rate, **ESRF**: European Synchrotron Radiation Facility; **ESI-MS**: electrospray-ionization mass spectrometry; **FCR**: fraction of charged residues; **HeV**: Hendra virus; **IDP/IDR**: intrinsically disordered protein/ intrinsically disordered regions; **IMAC**: immobilized-metal affinity chromatography; **MeV**: measles virus; **NCPR**: net charge per residue; **NF**: natively folded; **N_{TAIL}**: C-terminal domain of measles virus nucleoprotein; **Pro**: proline; **PNT4**: residues 300-404 of Hendra virus phosphoprotein; **PMG**: pre-molten globule; **R_S**: Stokes radius; **Rg**: gyration radius; **SEC**: size exclusion chromatography; **U**: unfolded



Highlights

- Conformation of IDPs responds to linear distribution of oppositely charged residues
- The degree of charge clusterization is described by the parameter "κ"
- Increasing κ values correspond to progressive charges clusterization
- Charge clusterization induces protein compaction
- High proline content reduces the compaction responsiveness to charge clusterization

Abstract

Intrinsically disordered proteins (IDPs) have a biased amino acid composition, being highly enriched in charged residues compared to globular proteins. Recent theoretical and computational studies have shown that the charge content and, most importantly, the linear distribution of opposite charges are major determinants of the conformational properties of IDPs. The charge segregation in a sequence can be measured through the parameter designated as κ, which represents a weighted mean squared deviation of charge asymmetry. A strong inverse correlation between κ and gyration *radii* has been previously demonstrated for two independent sets of permutated IDP sequences. To assess to which extent conclusions based on charge clustering can be generalized, we have used two well-characterized IDPs, namely measles virus N_{TAIL} and Hendra virus PNT4, sharing a very similar fraction of charged residues (FCR~0.3) and net charge per residue (|NCPR|<0.05), but differing in Pro content (~11% and ~5%, respectively). For each protein, we have rationally designed a low- and a high-κ variant endowed with the highest and the lowest κ values compatible with their natural amino acid composition. Then, the conformational properties of *wt* and κ-variants have been assessed by biochemical and biophysical techniques. As expected, experimental data show a direct correlation between κ value and protein compaction. Moreover, the extent of response to charge clustering suggests a critical role for the content of Pro residues. The abundance of this amino acid emerges as a main cause of resilience to conformational compaction governed by charge patterns. These results contribute to shed light onto the sequence determinants of conformational properties of IDPs and may also serve as an asset for the rational design of non-natural IDPs with a desired degree of compactness.

1. Introduction

Intrinsically disordered proteins/regions (IDPs/IDRs) exist as dynamic ensembles subjected to extreme fluctuations of atom positions and backbone Ramachandran angles (Dunker et al., 2001a; Habchi et al., 2014; Romero et al.; Tompa, 2012a; Uversky, 2002a, 2009; Uversky et al., 2000; Wright and Dyson, 1999). This behaviour is imparted by their peculiar amino acid composition. The sequence of IDPs/IDRs is indeed enriched in "disorder promoting" residues, such as Ala, charged residues and structure-breaking residues (Pro and Gly) and depleted in bulky (e.g., hydrophobic and aromatic) residues (Dunker and Obradovic, 2001; Dunker et al., 2008; Tompa, 2012a; Uversky, 2002a). Intrinsic disorder is abundant in nature, especially in the proteome of complex organisms (Dunker and Obradovic, 2001; Schad et al., 2011; Ward et al., 2004), where signalling and regulatory functions are also more present (Tompa, 2012a). Bioinformatics analyses support a general correlation between intrinsic disorder and various diseases, such as cancer, diabetes, amyloidoses, neurodegenerative and cardiovascular diseases (Uversky et al., 2008). Because of the many biological roles they play, the number of studies focused on IDPs has never ceased to grow in the last two decades. More recently, IDPs have gained momentum as the knowledge of the mechanisms underlying their conformational properties can help in designing synthetic polypeptides that can find very diverse applications, such as "entropic bristles" (Santner et al., 2012), "stimuli sensitive brushes" (Srinivasan et al., 2014), or novel biomaterial endowed with sol-gel properties (Simon et al., 2017b).

Despite the growing interest being paid to IDPs, the debate on determinants of their conformational properties is still lively. In recent years, research efforts from various groups have been focused on deciphering how the primary structure encodes IDP conformation and aggregation, highlighting the role of sequence length (Uversky et al., 2012), hydrophobic interactions (Felitsky et al., 2008;

Hiller et al., 2008), hydrogen bonds (Bolen and Rose, 2008; Möglich et al., 2006) and electrostatic charges (Gast et al., 1995; Hofmann et al., 2008; Mao et al., 2010; Tedeschi et al., 2017; Zhou, 2002).

Among these, elegant studies from the group of Rohit Pappu have shown that the net charge per residue (NCPR, defined as $|f_+ - f_-|$, where f_+ and f_- are the fractions of positively charged and negatively charged residues, respectively) (Mao et al., 2010), the total fraction of charged residues (FCR, defined as $f_+ + f_-$), and the linear distribution of opposite charges (Das and Pappu, 2013) are the primary determinants of the chain dimensions and conformational classes of IDPs (Das et al., 2015). The main results of these works are valuable predictive tools: the so-called Das-Pappu diagram of state (Mao et al., 2010), and the "κ" parameter (Das and Pappu, 2013; Das et al., 2015). The Das-Pappu diagram establishes a relationship between FCR values of proteins and various classes of structural conformation (Das et al., 2015). The κ parameter is the mean squared deviation of charge asymmetry weighted on the maximal asymmetry allowed for a given amino-acid composition (Das and Pappu, 2013). The value of κ is comprised between 0 (opposite charges evenly distributed) and 1 (opposite charges segregated into two clusters) and an inverse correlation has been demonstrated between κ and gyration radius (R_g) of IDPs and unfolded proteins (Das and Pappu, 2013; Das et al., 2015). Seminal computational studies have been nicely complemented by experimental studies carried out on ~100-residue natural IDRs and their permutants designed to sample various κ values. Specifically, κ-variants were obtained by maintaining unchanged the native NCPRs and amino acid compositions, and modifying the position of both charged and non-charged amino acids along the sequence (Das et al., 2016; Sherry et al., 2017).

Here, we have conceived κ variants by keeping unchanged all non-charged residues and permutating only the positions of charged ones in model proteins. The model systems we chose are two well-characterized IDPs, a 124-residue

region from the measles virus (MeV) nucleoprotein N (N_{TAIL}) (Longhi et al., 2017; Longhi et al., 2003) and a 104-residue region from the Hendra virus (HeV) phosphoprotein P (PNT4) (Habchi et al., 2010). They were chosen because of their different content in Pro (Pro ~11% and ~5%, for N_{TAIL} and PNT4, respectively), and of their rather high FCR (~0.3) and low |NCPR| values (<0.05). As already mentioned, Pro has been recognized as a "disorder promoting" amino acid (Dunker et al., 2001a) and a major determinant of compaction (Marsh and Forman-Kay, 2010). Indeed, Pro is an α- and β-structure breaker due to the conformational constraints imposed by its pyrrolidine ring (Adzhubei and Sternberg, 1993). Moreover, the tendency of Pro residues to undergo *cis-trans* isomerization may hamper the structural compaction of IDPs and control their size (i.e., their Stokes radius - R_S) (Marsh and Forman-Kay, 2010). Overall, high frequency of Pro among IDPs has been hypothesized to be evolutionary conserved, since related to their compactness and, hence, to their biological function (Marsh and Forman-Kay, 2010). Besides the Pro content, the fraction of charged residues, namely the FCR, dictates the size, shapes and amplitudes of conformational fluctuations of IDPs (Das and Pappu, 2013; Das et al., 2015; Konarev et al., 2003). On FCR depends the highest κ value can be obtained for a given sequence. Starting from the κ values of 0.159 and 0.213 for the *wt* N_{TAIL} and PNT4, respectively, the highest κ value obtained by sequence permutations were ~0.421 and 0.431, respectively. Low NCPR values of our model proteins reflect an equilibrated number of negatively and positively charged residues, which makes possible to reach rather low κ values by sequence permutation (0.044 and 0.078 for PNT4 and N_{TAIL}, respectively). Overall, we have produced two sets of proteins, each including a *wild-type* (*wt*), a low-*κ* and a high-*κ* variant. Their conformational properties have been investigated using various biophysical approaches. The results show that the proteins analysed here experience a conformational compaction that is directly related to charge clustering, and hence

to κ values, and inversely related to their Pro content.

2. Materials and methods

2.1 Gene design and cloning

The two model IDPs used in this work are PNT4, corresponding to the region encompassing residues 300-404 of the HeV P protein (Habchi et al., 2010), and N_{TAIL}, corresponding to the 401-525 region of MeV nucleoprotein (Bourhis et al., 2004; Longhi et al., 2003). For each model IDP, the *wt*, low-κ and high-κ variants share the same number of charged residues, but differ in the way they are distributed along the sequence. Note that the changes in distribution involve just the positions occupied by charged residues whereas the positions of non-charged residues are left unchanged in all variants. In low-κ sequences, positively (Lys and Arg) and negatively (Glu and Asp) charged residues are more evenly distributed than in the *wt* parental sequence. By contrast, in high-κ sequences, positively and negatively charged residues are clustered in the N- and C-terminal region, respectively. The κ parameter, along with NCPR and FCR, were calculated using the CIDER webserver (Holehouse et al., 2017). Synthetic genes optimized for expression in *Escherichia coli* and encoding the κ variants were obtained through gene synthesis (Genscript, Piscataway, NJ, USA), and were cloned into the pET-21a vector (EMD, Millipore, Billerica, MA, USA), between the *Nde*I and *Xho*I sites, giving rise to the plasmids pET-21a [PNT4 *wt*], pET-21a [PNT4 low-κ], pET-21a [PNT4 high-κ], pET-21a [N_{TAIL} *wt*], pET-21a [N_{TAIL} low-κ] and pET-21a [N_{TAIL} high-κ]. Each synthetic gene encodes a protein with an N-terminal hexa-histidine (6xHis) tag, while a stop codon is inserted immediately before the *Xho*I restriction site thereby excluding from the coding region the vector-encoded 6xHis tag. The sequence of all the constructs was checked by DNA sequencing (GATC-Biotech, Koeln, Germany) and found to conform to expectations. The

amino acid sequences are shown in **Figure 1.**

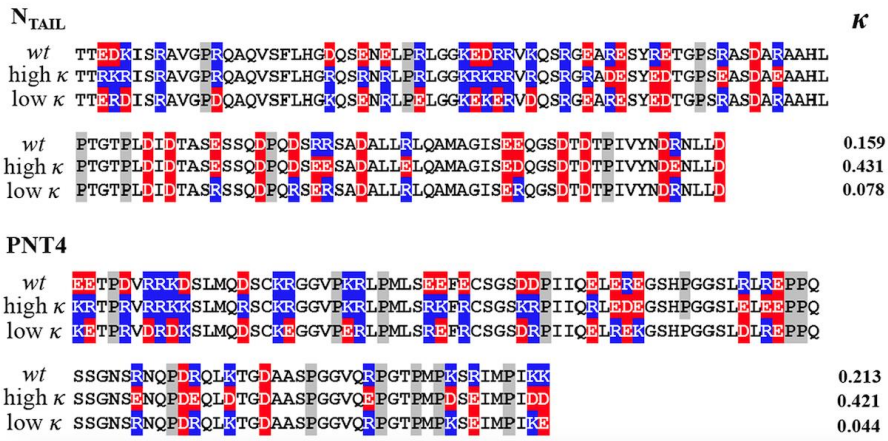


Figure 1. *κ* variants sequences. Sequences of *wt* and *κ* variants of N_{TAIL} and PNT4. Common regions inside each set of sequences (starting Met, 6xHis) are not shown. Residues of Pro are shown in grey, positively and negatively charged residues are highlighted in red and blue, respectively; names of sequences are indicated on the left and κ values on the right.

2.2 Protein expression and purification

The *E. coli* strain T7 pLysS (New England Biolabs, Ipswich, MA, USA) was used as the host for heterologous expression of all protein variants. Transformed cells were grown overnight at 37° C in Lennox medium (10 g/L tryptone, 5 g/L yeast extract, 5 g/L NaCl), then diluted 1 : 20 in 200 mL of Zym-5052 medium (Studier, 2005) and incubated for 15 hr at 25 °C under shaking at 220 rpm. All culture media were supplemented with 100 mg/L ampicillin. Cultures were harvested by centrifugation, and the cell pellets were frozen at -20°C. Cell pellets were resuspended in 5 volumes (v/w) of lysis buffer (50 mM NaH_2PO_4, 1 M NaCl, 10 mM imidazole) supplemented with EDTA-free protease inhibitor cocktail (Sigma-Aldrich, St. Louis, MO, USA), and disrupted by sonication using a Branson 450 Sonifier (Emerson Electric Co., St. Louis, MO, USA) (five cycles of

30 s each at 60% power output). Lysates were boiled for 5 minutes and clarified by centrifugation (15,000 *g* for 30 min). To increase the purification yields of high-κ variants and of *wt* PNT4, guanidium chloride was added to the lysate to a final concentration of 6 M and the lysate was directly clarified by centrifugation as described above omitting the boiling step.

Taking advantage of the presence of the 6xHis tag, all the proteins used in this study were purified by immobilized metal affinity chromatography (IMAC) using a fast protein liquid chromatography (FPLC) Äkta system (GE, Healthcare, Little Chalfont, UK) equipped with a His-Trap HP column (GE Healthcare, Little Chalfont, UK). The elution fractions containing the highest protein concentration were pooled and subjected to size exclusion chromatography (SEC) using the same Äkta system and a Superdex 75 16/60 column (GE, Healthcare, Little Chalfont, UK). The column was equilibrated in SEC buffer (10 mM ammonium acetate pH 7.0, 2 mM EDTA, 5% glycerol, supplemented with 500 mM NaCl for high-κ variants) and the protein of interest was eluted in SEC buffer with an elution rate of 1 mL/min. Protein concentration of N_{TAIL} variants was determined at 280 nm using a Nanodrop spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) and the theoretical absorption coefficient as obtained from the Expasy server (https://www.expasy.org/). In the case of PNT4 variants that lack Tyr and Trp residues, protein concentration was estimated using a Bradford protein colorimetric assay (Bio-Rad Laboratories Inc., Hercules, CA, USA) and bovine serum albumin as a standard.

2.3 Protease sensitivity assays

Limited trypsin proteolysis (Promega Corp., Madison, WI, USA) of purified N_{TAIL} and PNT4 variants (1 mg/mL) was performed in 10 mM Tris/HCl buffer at pH 7.8, at room temperature. 0.25 µg of trypsin were used per 1 µg of target protein, in a total volume of 300 µL. The extent of proteolysis was analysed by 18% SDS-

PAGE of 9-µg samples withdrawn from the reaction mixture at different times (0, 2, 6, 8, 10 and 12 hours) after the reaction start. The reaction was stopped by mixing each sample with 10 µL 2x Laemmli sample buffer and boiling for 5 minutes.

2.4 Far-UV circular dichroism (CD) spectroscopy

Far-UV CD analyses were carried out on samples in ammonium acetate. Circular dichroism spectra were recorded with a Jasco J-810 spectropolarimeter (Jasco Corp., Easton, MD, USA), in a 1-mm path-length quartz cuvette. Measurements were performed at variable wavelengths (195–260 nm) with scanning velocity of 20 nm/min and data pitch of 0.2 nm. All spectra were corrected for buffer contribution, averaged from two independent acquisitions, and smoothed by the Means-Movement algorithm implemented in the Spectra Manager package. Experiments were performed in triplicate. Mean ellipticity values per residue ([θ]) were calculated as:

$$[\theta] = \frac{3300 \cdot m \cdot \Delta A}{c \cdot n \cdot l} \qquad \text{(Eq. 1)}$$

where ΔA is the difference in the absorption between circularly polarized right and left light of the protein corrected for blank, l is the path length (in cm), n is the number of residues, m is the molecular mass (in Daltons), and c is the protein concentration (in mg/mL). For all samples, the concentration was 0.2 mg/mL. The number of residues of N_{TAIL} and PNT4 are reported in **Table 2.**

2.5 Hydrodynamic radius estimation by SEC

The hydrodynamic radius or Stokes radius (R_S) of proteins composed by N amino acids were calculated from the following equations describing the behaviour of a set of natively folded (R_S^{NF}), chemically denatured (R_S^{U}), and his-tagged, intrinsically disordered (R_S^{IDP}) proteins made of N residues (Uversky, 2002b):

$$R_S^{NF} = 4.92\ N^{0.285} \quad \text{(Eq. 2)}$$

$$R_S^{U} = 2.33\ N^{0.549} \quad \text{(Eq. 3)}$$

$$R_S^{IDP} = (AP_{pro} + B)(C|Q| + D)S_{his}R_0N^{\nu} \quad \text{(Eq. 4)}$$

where P_{pro} = fraction of prolines, |Q|= number of glutamines, A = 1.24; B = 0.904, C = 0.00759, D = 0.963, S_{his} = 0.901, R_0 = 2.49 and ν = 0.509.

To compare the degree of compaction in a way independent of N, we have calculated the compaction index CI_{Rs} (Brocca et al., 2011; Wilkins et al., 1999):

$$CI_{Rs} = \frac{R_S^U - R_S^{exp}}{R_S^U - R_S^{NF}} \quad \text{(Eq. 5)}$$

where R_S^{exp} is the experimental value for a given protein, R_S^{U} and R_S^{NF} are the reference values calculated for an unfolded protein, according to equation 3, and for a globular folded protein, according to equation 2, respectively.

2.6 Dynamic Light Scattering (DLS) studies

Dynamic light scattering experiments were performed using a Zetasizer Nano-S (Malvern Instruments, Malvern, Worcestershire, UK) at 25°C. Samples were at 1 mg/mL in SEC buffer supplemented with 500 mM NaCl for high-κ variants. The R_S was deduced from the translational diffusion coefficient using the Stokes-

Einstein equation. Diffusion coefficients were inferred from the analysis of the decay of the scattered intensity autocorrelation function. All calculations were performed using the software provided by the manufacturer.

2.7 Small Angle X-ray scattering (SAXS) studies

All small-angle X-ray scattering (SAXS) measurements were carried out at the European Synchrotron Radiation Facility (Grenoble, France) on beamline BM29 (bending magnet) at a working energy of 12.5 KeV. Data were collected on a Pilatus (1M) detector. The wavelength was 0.992 Å. The sample-to-detector distance of the X-rays was 2.847 m, leading to scattering vectors q ranging from 0.028 to 4.525 nm^{-1}. The scattering vector is defined as $q = 4\pi/\lambda\sin\theta$, where 2θ is the scattering angle. The exposure time was optimized to reduce radiation damage. SAXS data were collected at 20 °C using purified protein samples (50 µL each). Proteins were analysed at various concentrations. The ranges of concentration used in these studies differed from protein to protein reflecting their different aggregation propensities. Samples were in SEC buffer supplemented with 300 mM arginine in the case of N_{TAIL} *wt*, N_{TAIL} low-κ and PNT4 low-κ (and with 500 mM NaCl in the case of high-κ variants). Samples were loaded in a fully-automated sample charger. Ten exposures of 10 s each were made for each protein concentration and data were combined to give the average scattering curve for each measurement. Data points affected by aggregation, possibly induced by radiation damages, were excluded. For all the PNT4 variants, which are more prone to aggregation, we merged the scattering curves to exclude data possibly affected by aggregation, whereas for N_{TAIL} variants, which are comparatively less affected by aggregation, we used the scattering curves at the highest concentration to obtain maximal information at high resolution.

The data were analyzed using the ATSAS program package (Petoukhov et al., 2012). Data reduction was performed using the established procedure available at

BM29, and buffer background runs were subtracted from sample runs. The R_g and forward intensity at zero angle I(0) were determined with the program PRIMUS (Konarev et al., 2003), according to the Guinier approximation at low q values, in a $q \cdot R_g$ range up to 1.3.

$$Ln[I(q)] = Ln[I_0] - \frac{q^2 R_g^2}{3} \quad \text{(Eq. 6)}$$

The forward scattering intensities were calibrated using water as reference. The R_g and pair distance distribution function, P(r), were calculated with the program GNOM (Svergun, 1992). The maximum dimension (D_{max}) value was adjusted such that the R_g value obtained from GNOM agreed with that obtained from the Guinier analysis.

For each protein, we also attempted at describing it as a conformational ensemble. To this end we used the program suite EOM 2.0 (Tria et al., 2015) using the default parameters.

The theoretical values of R_g (in Å) of proteins composed by N amino acids were calculated from the following equations describing the behaviour of a set of natively folded (R_g^{NF}) (Wilkins et al., 1999), chemically denatured (R_g^{U}) (Tria et al., 2015), and an intrinsically disordered (R_S^{IDP}) (Wilkins et al., 1999) proteins of known length:

$$R_g^{NF} = \sqrt{(3/5)}\ 4.75\ N^{0.29} \quad \text{(Eq. 7)}$$

$$\text{Log}\ (R_g^{U}) = 0.58\ \text{Log}\ (N) + 0.80 \quad \text{(Eq. 8)}$$

$$R_g^{IDP} = R_0 \cdot N^{\nu} \quad \text{(Eq. 9)}$$

where R_0 is 2.54 ± 0.01 and ν is 0.522 ± 0.01.

By analogy with what we did with R_S, we calculated an R_g-based compaction index CI_{Rg} (Brocca et al., 2011; Wilkins et al., 1999):

$$CI_{Rg} = \frac{R_g^U - R_g^{exp}}{R_g^U - R_g^{NF}} \quad \text{(Eq. 10)}$$

where R_g^{exp} is the experimental value for a given protein, R_g^U and R_g^{NF} are the reference values calculated for an unfolded protein according to equation 8 and for a globular folded protein according to equation 7, respectively.

2.8 Electrospray-ionization mass spectrometry (ESI-MS)

An aliquot of each sample obtained by SEC was diluted in SEC buffer to a final protein concentration of 20 µM, and 10 µL of the final solution were directly injected into the spectrometer under denaturing conditions, employing borosilicate-coated capillaries of 1 µm internal diameter (Thermo Fisher Scientific, Waltham, MA, USA). Nano ESI-MS spectra were acquired in positive-ion mode on a hybrid quadrupole-time-of-flight spectrometer (QSTAR Elite, AB Sciex, Foster City, CA, USA). The main instrumental parameters were: ion spray 1.1 kV; declustery potential 60 V; curtain gas 20 PSI. Final spectra were averaged over 1-min acquisition time.

3. Results

3.1 Design of κ variants for N_{TAIL} and PNT4

The model IDPs used in this study are N_{TAIL} from MeV and PNT4 from HeV. These proteins have lengths and Pro content similar to those of IDPs already used to experimentally demonstrate the influence of charge patterning on IDP conformation (Das et al., 2016; Sherry et al., 2017). The rationale for choosing these two IDPs is the following. Both *wt* proteins have a pre-molten globule (PMG) conformation and are located in the region 2 (R2) of the Das-Pappu state diagram (data not shown) (Das and Pappu, 2013; Mao et al., 2010). The conditions for the assignment of an IDP sequence to this region are the FCR value (0.25<FCR<035), its length and Pro content, which should be "reasonably low" (Das et al., 2015). Proteins belonging to R2 (0.25< FCR<0.35) are mainly "Janus sequences", which are collapsed or expanded in a context-dependent manner, are statistically the most abundant (Das et al., 2015). This type of proteins are more responsive not only to environment changes (e.g., salt concentration, pH, ligand binding etc), but also to changes in primary structure (Das and Pappu, 2013; Das et al., 2015; Mao et al., 2010). This makes us more confident in inducing detectable compactness changes through changes of charges patterning. The FCR of N_{TAIL} and PNT4 (0.299 and 0.298) is rather high and dictates the highest κ value attainable by sequence permutation. On the contrary, the low NCPR values of the two model proteins (-0.045 for N_{TAIL} and 0.018 for PNT4) reflect an overall balanced number of positively and negatively charged residues (see below), and allows minimizing the κ value of permutants.

For each model IDP, two *"κ variants"* were designed by keeping unchanged the amino acid composition of the *wt* protein. The resulting κ variants thus differ solely in the patterning of charged residues (Arg, Lys, Asp and Glu). Noteworthy, in our design, non-charged residues maintain their original position, while positive

and negative charges are clustered in two blocks (high-*κ* variants), or distributed as much evenly as possible along the sequence (low-*κ* variants). In the case of N_{TAIL}, starting from the *wt* κ =0.159, sequence permutation brought to *κ* values of 0.078 and 0.431 (Δκ=0.353). In the case of PNT4, starting from the *wt* κ = 0.213, we have obtained *κ* values of 0.044 and 0.421 (Δκ=0.377). These represent the highest and the lowest *κ* values compatible with their amino acid composition and with the constraints used for the sequence design (**Figure 1)**. The *κ* values obtained for the two model proteins are rather similar due to the similarity of their FCR.

The disorder profiles of all the proteins, as obtained using Pondr-fit (Xue et al., 2010a), are shown in **Figure 2B** and **2C**. Overall, permutants exhibit a rather disordered profile, with high-*κ* variants reaching the highest disorder scores. PNT4 variants are predicted to be more disordered than N_{TAIL} variants although endowed with the lower content of P.

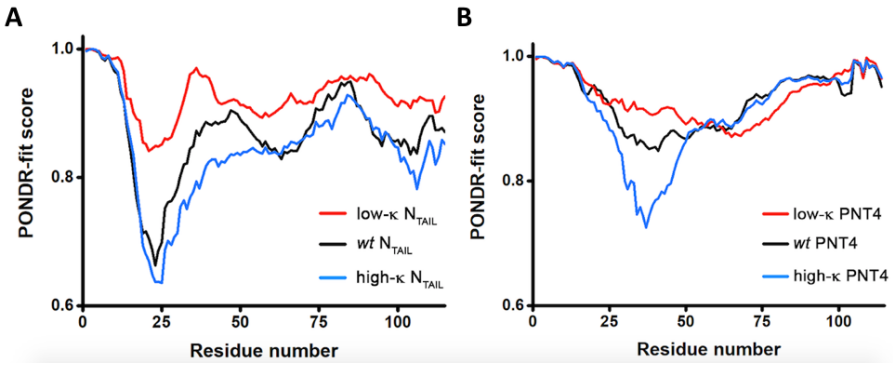


Figure 2. *κ* variants disorder predictions. **A**) Disorder prediction of N_{TAIL} and **B**) disorder prediction of PNT4 variants carried out using the Pondr-fit predictor.

3.2 Expression and purification of κ variants

All the proteins were purified from the soluble fraction of *E. coli* through IMAC and SEC. The high-κ variants displayed a high aggregation propensity. Addition of 500 mM NaCl prevented aggregation in the concentration range used in this study. PNT4 high-κ was the variant with the highest aggregation propensity. Typical purification yields from 1 L of bacterial culture for N_{TAIL} and PNT4 variants were ~ 5 mg of protein for low-*κ* and *wt* variants, and ~ 3 mg of protein for high-*κ* variants. Their purity was assessed by SDS-PAGE (data not shown) and the protein mass confirmed by mass spectrometry analyses. It should be noted that permutants of the same protein exhibit a different electrophoretic mobility, overall lower for high-*κ* variants. These differences are likely due to the clustering of positively charged residues, which may hinder their interaction with the sulphate groups of the detergent (Reynolds and Tanford, 1970).

3.3 Different protease sensitivity of κ variants

As a first step towards the assessment of the conformational properties of the *κ* variants, we carried out limited proteolysis experiments. The latter allow probing the overall solvent exposure and flexibility, which constitutes a hallmark of structural disorder (Receveur-Bréchot et al., 2006).

As shown in **Figure 3A**, *wt* N_{TAIL} and its high-*κ* variant show a very moderate degradation under the limited-proteolysis conditions employed here. In both cases, the highest band corresponding to the entire protein remains unvaried even after a 12-hour incubation with trypsin. In the case of the *wt* protein, the band pattern does not change during the time course of kinetics, although some proteolysis seems to have occurred even prior the incubation with trypsin. By contrast, the digestion of the N_{TAIL} low-*κ* variant follows a faster kinetics, with the full-length protein being no more detectable already after 8 hours of incubation (**Figure 3A**).

In the case of PNT4, the *wt* variant undergoes a pronounced degradation at early time of trypsin digestion, with the full-length protein being no more detectable after 8 hours of trypsin incubation. Note the accumulation of a partial digestion product of ~14 kDa. On the other hand, the high-*κ* variant displays a degradation pattern markedly different from that observed for *wt* PNT4. Besides the fact that it appears partially digested since the beginning of the kinetics, the full-length protein (see band of an apparent molecular mass of ~ 20 kDa) progressively vanishes as a function of time without however completely disappearing. At the same time, a fragment of apparently ~16 kDa appears early and persists during the whole time course. Conversely, low-*κ* PNT4 is completely degraded just after 1 hour of incubation, indicating the higher susceptibility of this variant compared to the others (**Figure 3B**).

Taken together, these results clearly indicate that the low-*κ* variants of the two proteins are more susceptible to trypsin degradation than high-*κ* and *wt* proteins. This behaviour can be ascribed either to the higher compaction of high-*κ* and *wt* variants with respect to low-*κ* ones, or to a different patterning of proteolytic sites arising from the different distribution of Lys and Arg residues. The latter hypothesis seems however unlikely as judged from the fact that the low-*κ* and *wt* variants, although sharing very similar primary structure and distribution of basic residues, present different proteolytic profiles.

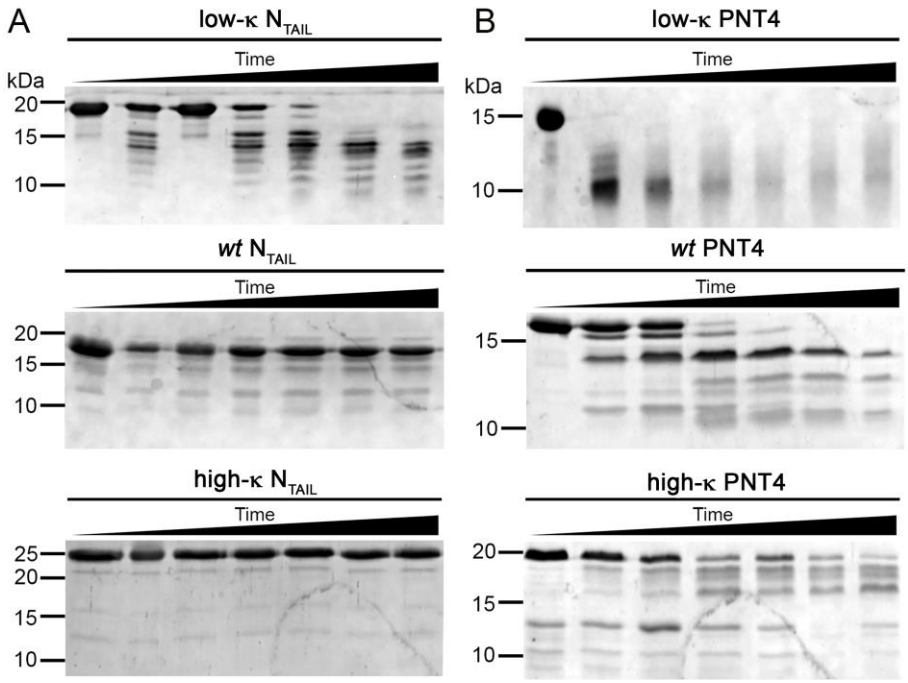


Figure 3. Trypsin sensitivity assay. SDS-PAGE analysis of the extent of digestion of purified proteins at different times (0, 2 h, 4 h, 6 h, 8 h, 10 h and 12 h) of incubation with trypsin. **A**) N_{TAIL} and **B**) PNT4.

3.4 Impact of charge pattern on secondary structure studied by CD spectroscopy

Far-UV CD spectroscopy analyses were carried out to obtain information on the secondary structure content of the N_{TAIL} and PNT4 variants (**Figure 4**).

The spectrum of *wt* N_{TAIL} displays a pronounced negative peak at 198 nm and is superimposable to that already published for this protein (Longhi et al., 2003) (**Figure 4A**). Besides this typical trait of structural disorder, *wt* N_{TAIL} shows a small shoulder centred around 222 nm. The overall profile indicates the existence of some seeds of secondary structure, typical of the PMG state (Uversky, 2002a). This is likely due to the contribution of a partly pre-configured α-helix within a Molecular Recognition Element (MoRE) encompassing residues 89-106 (Habchi

and Longhi, 2012). The high-κ N_{TAIL} exhibits an even less disordered profile. The spectrum of low-κ N_{TAIL} has a flatter profile for wavelength higher than 210 nm, thus appearing slightly more disordered than its *wt* and high-κ counterparts.

As for *wt* N_{TAIL}, the CD spectrum of *wt* PNT4 shows typical traits of structural disorder mixed with some elements of secondary structure. Spectra of *wt* and low-κ PNT4 are almost superimposable, with a negative peak at 198 nm. Similarly to N_{TAIL}, the high-κ variant of PNT4 shows a shift from 198 nm to 205 nm (**Figure 4B**).

Althought endowed with the lower content of Pro, PNT4 variants are more disordered than N_{TAIL} ones, in agreement with Pondr-Fit profiles, and overall less sensitive to changes of charge patterning. Among N_{TAIL} permutants, the secondary structure content decrease in low-κ and increase in high-κ, compared to *wt*. Overall, the increase in secondary structure content correlates with charge clustering and was not predicted by Pondr-Fit.

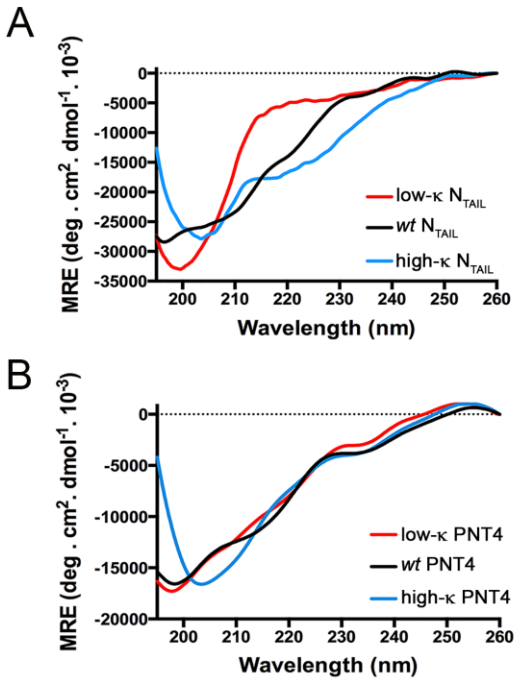


Figure 4. Far-UV CD spectra. Proteins were in SEC buffer at 0.2 mg/mL. Spectra were recorded at 20°C and were the means of two acquisitions. Shown are spectra representative of one out of three independent experiments. **A**) N_{TAIL} and **B**) PNT4.

3.5 Effects of charge pattern on hydrodynamic radius

To gain insights into the degree of compactness of N_{TAIL} and PNT4 variants, they were subjected to size exclusion chromatography (SEC) and dynamic light scattering (DLS). The two techniques gave comparable results (data not shown) and in the following we will comment only the results obtained through SEC. The R_S obtained for *wt* N_{TAIL} (24.1 Å ± 0.1) is close to that described in (Longhi et al., 2003). **Table 1** summarizes experimental results and theoretical data calculated with equation 2- 4 (Uversky, 2002b) for IDPs of the same length and with the same P and Q content (R_S^{IDP}). Indeed, the experimental value, 24.1 ± 0.1 Å, has to be compared with the theoretical 26.9 Å for N_{TAIL}. An even larger difference was observed between the experimental 22.5 ± 0.5 Å and the theoretical 26.3 Å of

PNT4. Overall, low-κ variants of both model proteins are slightly more extended than their respective *wt* counterpart, whereas high-κ variants have a smaller hydrodynamic radius than the *wt* form (**Table 1**).

Protein	Rs^{NF}	Rs^{U}	Rs^{IDP}	Variant	Rs^{SEC}	CI_{Rs}
N_{TAIL}	19.9	34.3	26.9	*wt*	24.1 ± 0.1	0.71
				high-κ	21.6 ± 0.2	0.88
				low-κ	24.8 ± 0.2	0.66
PNT4	19.0	31.4	26.3	*wt*	22.5 ± 0.5	0.72
				high-κ	21.8 ± 0.4	0.77
				low-κ	23.4 ± 0.1	0.64

Table 1. Stokes *radii* (R_s) in Å. The table shows the Stokes *radii* experimentally estimated by SEC (R_s^{SEC}), along with those calculated for natively folded (R_s^{NF}, from Eq. 2), chemically denatured (R_s^{U}, from Eq. 3), and intrinsically disordered proteins (R_s^{IDP}, from Eq. 4). The R_S-based compaction index (CI_{Rs}, from Eq. 5) is also shown. Shown are means and standard deviations from three independent experiments.

A useful parameter to compare the compactness of proteins with different number of residues (N) is the compaction index (*CI*) (Brocca et al., 2011; Wilkins et al., 1999). We referred to CI_{Rs} as a *CI* calculated using theoretical and experimental R_S values (see Eq. 5 in Materials and Methods). The comparison of the CI_{Rs} pinpoints that N_{TAIL} variants are more responsive than PNT4 variants to the clustering of electrostatic charges (**Table 1**). Indeed, the difference of CI_{Rs} (ΔCI_{Rs}) between the low- and the high-κ variants is 0.22 (0.88-0.66) in the case of N_{TAIL}, and 0.13 (0.77-0.64) in the case of PNT4. We have considered the regression of CI_{Rs} on the κ value. The relationship between the CI_{Rs} and κ value can be represented by the following linear functions:

N_{TAIL} CI_{Rs} = 0.631 κ + 0.608, with an R^2 = 0.999 (Eq. 11)

PNT4 CI_{Rs} = 0.340 κ + 0.634, with an R^2 = 0.991 (Eq. 12)

The slope of these functions, referred to as $^{\kappa}CR_{Rs}$, represents the compaction responsiveness to κ pertaining R_S (**Table 2, Figure 9A**). Overall, these data suggest that charge clustering promotes protein compactness to different extents in the two model proteins.

Protein	N	Proline (%)	Hydrophobicity	NCPR	FCR	$^{\kappa}CR_{Rg}$	$^{\kappa}CR_{Rs}$	Ref
N_{TAIL}	134*	11.4	3.35	-0.045	0.299	0.636	0.631	This work
PNT4	114*	5.2	3.24	0.018	0.298	0.988	0.340	This work
$p27_{96-198}$	108	9.3	3.26	0.009	0.254	0.647[a]	-	[29]
NAM	106	4.7	3.10	-0.028	0.368	-	1.561[b]	[30]

Table 2. Characteristics of IDPs analyzed for their responsiveness to κ changes. [a]) R^2 = 0.950; [b]) R^2 = 0.737.

*= N includes the starting Met and the 6xHis tag

3.6 Effects of charge pattern on conformation as inferred from ESI-MS studies

The conformational properties of the different variants were also investigated by non-denaturing ESI-MS, exploiting the dependence of the charge-states acquired during the electrospray on protein compactness [50. This method is particularly useful in the characterization of IDP conformational ensembles, thanks to its ability to distinctly detect protein conformers, even if present in a minor fraction of the molecular population (Natalello et al., 2017). Because of the strong aggregation propensity of PNT4, these proteins were excluded from the analyses. The spectrum of *wt* N_{TAIL} (**Figure 5B**) is characterized by broad peak-distribution at high-charge states and by a minor distribution (2%) centred on the 9+ ion state. This bimodal distribution is typical of the spectra of highly disordered, fluctuating

proteins. Low-*κ* N_{TAIL} (**Figure 5A**) shows an ESI-MS spectrum almost identical to the one of the low-*κ* variant, indicating that the two proteins present a very similar conformer population. On the other hand, the spectrum of high-*κ* variant (**Figure 5C**) is characterized by the increase of the component centred on the 9+ ion, that is 7-fold higher than in the case of the other two variants. The appearance of this compact form occurs at the expense of an intermediate component in the transition region, while the broad envelope of higher charges is only slightly modified (**Figure 5C**). A plausible explanation is that the high-*κ* N_{TAIL} exhibits an ensemble of multiple folded conformations in equilibrium with a more heterogeneous ensemble of more extended and heterogeneous ensemble. Therefore, ESI-MS experiments confirm the disordered nature of N_{TAIL} and its responsiveness to charge clustering.

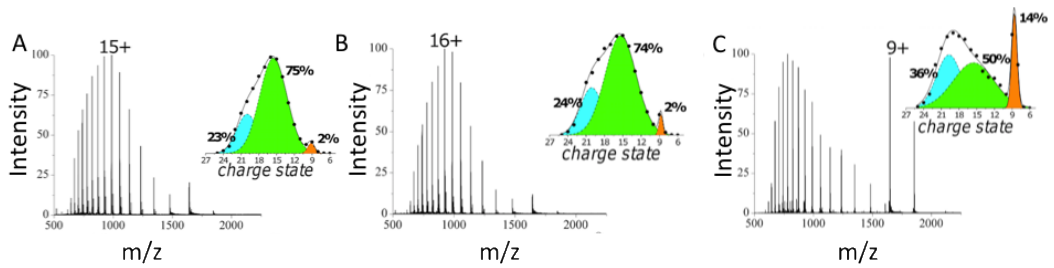


Figure 5. Native ESI-MS analyses. Mass spectra of low-κ **(A)**, *wt* **(B)**, and high-κ N_{TAIL} **(C)**. The most intense peaks of the spectra are labeled by the corresponding charge state. The inset in each panel represents the Gaussian fit of the charge-state distribution, where each Gaussian component is labeled by its relative amount.

3.6 Effects of charge pattern on the radius of gyration

To further analyse the effects of charge redistribution on protein compactness, we carried out SAXS analyses. For all variants, the shapes of the SAXS curves were independent of protein concentration (**Figures 6** and **7**) indicating the absence of significant aggregation under the experimental conditions used in these experiments. The Guinier plots, as obtained from the scattering curve either at the highest protein concentration (N_{TAIL} variants) or from merged data (PNT4 variants), are shown in **Figures 6A** and **7A**. Each curve can be well approximated by a straight line in the Guinier region ($q \cdot R_g$ <1.0). The slope of the curve is proportional to the R_g, while the intercept of the straight line gives the I(0) that is proportional to the molecular mass of the scatterer. For all the variants, the R_g values calculated by Guinier plots are in good agreement with the values determined from the pair distribution function P(r) (**Table 3** and **Figures 6B** and **7B**). For all the variants the molecular masses, as inferred from I(0) are in rather good agreement with the theoretical ones (**Table 3**).

For all the variants, the R_g and D_{max} values obtained at the various concentrations (see **Tables 4** and **5**) are in quite good agreement with each other's. Overall, comparison of the R_g and D_{max} values among κ variants reveals that both values decrease at increasing κ values (see **Tables 3** and **4**, **5),** indicating that the protein becomes more compact with increasing charge clustering.

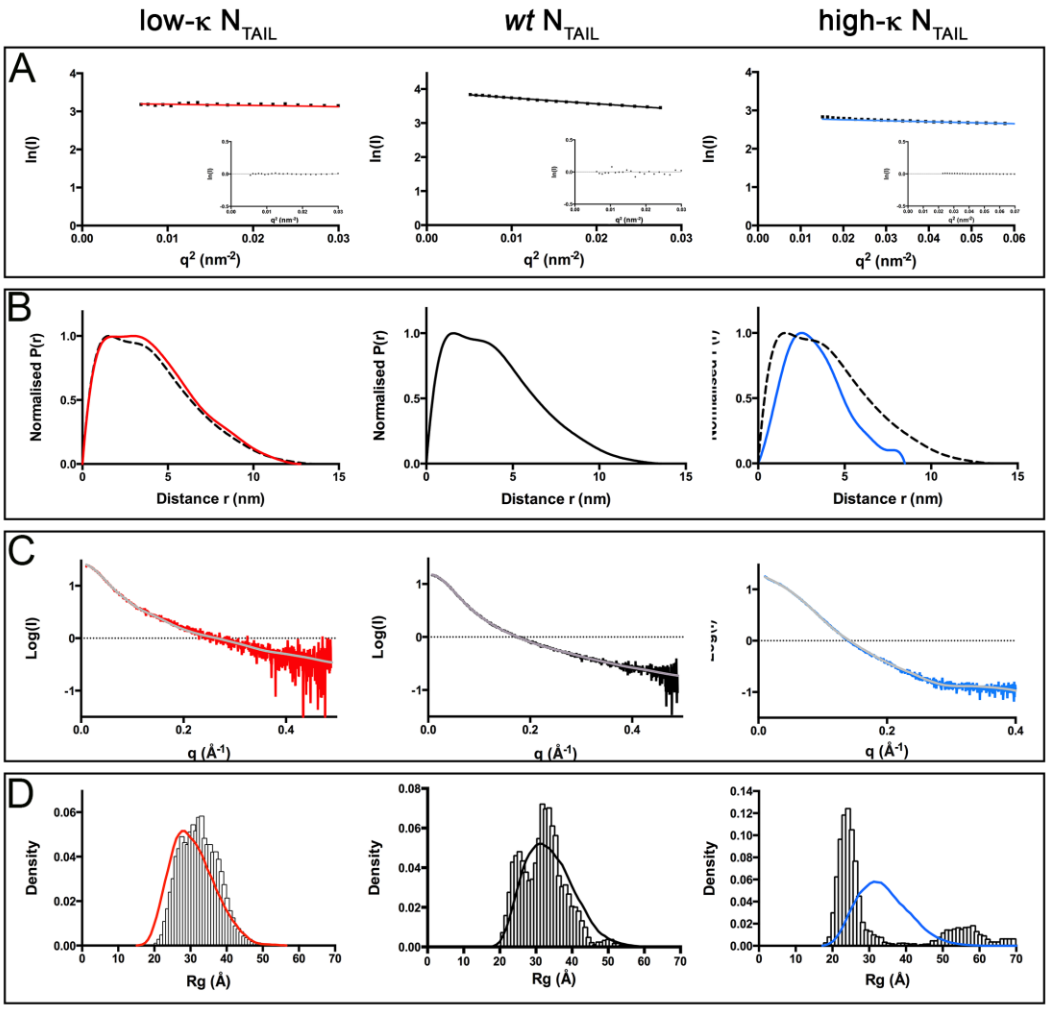


Figure 6. Small-angle X-ray scattering experiments of N_{TAIL} variants. A) Guinier plots obtained at 4.0 (*wt* N_{TAIL}), 3.2 (low-*κ* N_{TAIL}) and 5.8 (high-*κ* N_{TAIL}) mg/mL. Inset: residuals. **B)** Pair distance distribution, P(r), function of the data obtained at the same concentrations as in panel A. **C)** Experimental scattering curve of the proteins at the same concentrations as in panel A and EOM fit (grey) as obtained using Crysol. **D)** R_g distribution of the initial ensemble of randomly generated conformers (continuous line) and of the final sub-ensemble of selected conformers as obtained using EOM (bars).

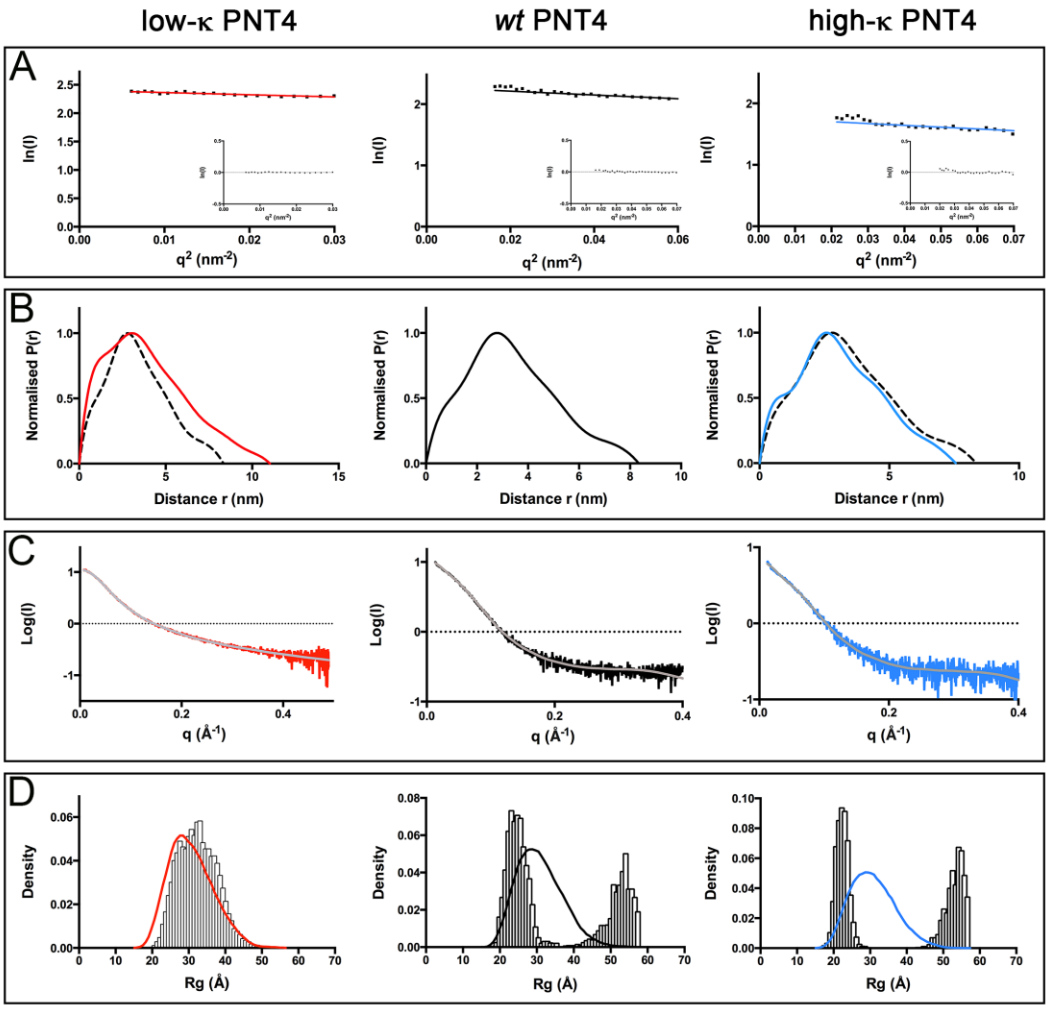


Figure 7. Small-angle X-ray scattering experiments of PNT4 variants. **A)** Guinier plots obtained from the merged curve obtained by merging the data at the various concentrations. Inset: residuals. **B)** Pair distance distribution, P(r), function of the merged data. **C)** Experimental merged scattering curve of the proteins and EOM fit (grey) as obtained using Crysol. **D)** R_g distribution of the initial ensemble of randomly generated conformers (continuous line) and of the final sub-ensemble of selected conformers as obtained using EOM (bars).

	wt N_{TAIL}	high-κ N_{TAIL}	low-κ N_{TAIL}	*wt* PNT4	high-κ PNT4	low-κ PNT4
Data collection parameters						
Concentration range (g/L)	1.2 – 4.0	0.6 - 2.0	0.6 -3.2	1 – 6.2	1 – 4.5	3.0 - 5.0
Structural parameters						
I(0) (cm^{-1}) (from Guinier)	15.0 ± 0.01	15.9 ± 0.01	17.6 0.01	±11.4 0.07	±13.3 0.03	±10.9 ± 0.02
R_g (Å) (from P (r))	31.0	26.0	32.0	28.0	24.4	33.0
R_g (Å) (from Guinier)	30.0 ± 0.1	26.0 ± 0.1	31.0 0.1	±28.4 0.2	±24.6 ± 0.1	32.0 ± 0.4
D_{max} (nm)	9.9	8.5	12.7	9.3	7.6	12.6
Molecular mass determination (kDa)						
Molecular mass (Da) (from I(0))	15000	15900	17650	11370	13350	10900
Calculated molecular mass (Da) from sequence	14775	14775	14775	12634	12634	12634

Table 3. SAXS data collection and scattering-derived structural parameters as obtained from either merged data (PNT4 variants) or data at the highest concentration (N_{TAIL} variants).

low-κ N_{TAIL} concentration (g/L)	R_g (Å) Guinier	D_{max} (Å)
0.6	31.0 ± 0.1	122
2.1	32.0 ± 0.4	114
3.2	31.0 ± 0.1	127

wt N_{TAIL} concentration (g/L)	R_g (Å) Guinier	D_{max} (Å)
1.2	30.0 ± 0.1	116
3.0	31.0 ± 0.2	124
4.0	30.0 ± 0.1	99

high-κ N_{TAIL} concentration (g/L)	R_g (Å) Guinier	D_{max} (Å)
1.0	25.8 ± 0.2	76
2.2	25.8 ± 0.1	78
5.9	26.0 ± 0.1	84

Table 4. R_g (from Guinier) and D_{max} for *wt*, high-κ, low-κ N_{TAIL} at various protein concentrations.

wt PNT4 concentration (g/L)	R_g (Å) Guinier	D_{max} (Å)
1.0	26.5 ± 0.3	88
2.9	28.4 ± 0.2	89
6.2	28.9 ± 0.2	93

low-κ PNT4 concentration (g/L)	R_g (Å) Guinier	D_{max} (Å)
3.0	31.0 ± 0.3	100
4.0	32.0 ± 0.1	128
5.0	32.0 ± 0.4	126

high-κ PNT4 concentration (g/L)	R_g (Å) Guinier	D_{max} (Å)
1.0	24.6 ± 0.1	78
2.6	25.2 ± 0.3	72
4.5	25.6 ± 0.3	76

Table 5. R_g (from Guinier) and D_{max} for *wt*, high-κ and low-κ PNT4 at various protein concentrations.

The experimental values of R_g are collected in **Table 6** where they are also compared to the theoretical R_g values expected for globular, fully unfolded and intrinsically disordered proteins of the same size (Wilkins et al., 1999) and calculated with equations 7-9. The experimental value of R_g obtained for *wt* N_{TAIL} is in good agreement with the previously published one (Longhi et al., 2003). For the *wt* form of both model proteins, the R_g values are slightly lower than expected for IDPs of identical length. For both model proteins, the experimental R_g values of the high-κ and low-κ variants are respectively smaller and either moderately (N_{TAIL}) or significantly (PNT4) larger than the value observed for the *wt* form (**Table 6**).

Protein	R_g^{NF}	R_g^{U}	R_g^{IDP}	Variant	R_g^{SAXS}	CI_{Rg}
N_{TAIL}	15.2	38.1	32.7	*wt*	30.0 ± 0.1	0.34
				high-κ	26.0 ± 0.1	0.52
				low-κ	31.0 ± 0.6	0.30
PNT4	14.4	34.7	30.1	*wt*	28.4 ± 0.2	0.30
				high-κ	24.6 ± 0.1	0.49
				low-κ	32.0 ± 0.7	0.12

Table 6. Gyration *radii* (R_g) in Å. The table shows the gyration radius experimentally measured by SAXS (R_g^{SAXS}), and those calculated for natively folded (R_g^{NF}, from Eq. 7), chemically denatured (R_g^{U}, from Eq. 8), and intrinsically disordered proteins (R_g^{IDP}, from Eq. 9). It also shows the R_g-based compaction index (CI_{Rg}, from Eq. 10).

A similar trend is observed for the D_{max} values (**Table 3**). Therefore, data support an increase in compactness with increasing charge partitioning. As we did in the case of hydrodynamic *radii*, we calculated CI_{Rg} from theoretical and experimental R_g values (see Eq. 10), and considered the regression of CI_{Rg} on κ value (**Table 2,**

Figure 9C). The relationship between the two parameters can be represented by the following linear functions:

N_{TAIL} CI_{Rg} = 0.636 κ + 0.254, with an R^2 = 0,998 (Eq. 13)

PNT4 CI_{Rg} = 0.988 κ + 0.079, with an R^2 = 0,996 (Eq. 14)

Hence, in contrast with data derived from experimental R_S, PNT4 variants are more responsive than N_{TAIL} variants to changes in κ values as far as their R_g is concerned.

We next analysed the dimensionless Kratky plots of the variants and compared them to the plots of a disordered, partially folded and folded protein (**Figure 8**). For both N_{TAIL} and PNT4 variants, the dimensionless Kratky plots reveal an overall gain of content in ordered structure with increasing κ, although *wt* PNT4 and high-κ PNT4 are less dissimilar from each other than their N_{TAIL} counterparts (see **Figure 8B** and **C**).

To further illuminate the dynamic behavior of the N_{TAIL} and PNT4 variants, we investigated the R_g distribution of the proteins using EOM (see Materials and Methods). From an initial pool of 10,000 random conformations, EOM selects a sub-ensemble of conformers that collectively reproduces the experimental SAXS data and represents the distribution of structures adopted by the protein in solution. The average SAXS scattering curves back-calculated from the selected sub-ensembles reproduce correctly the experimental curves (**Figures 6C** and **7C**).

The R_g distribution of the selected sub-ensemble of low-κ variants is symmetrical, as is that of *wt* N_{TAIL} (**Figures 6D** and **7D**). By contrast, the R_g distribution of the selected sub-ensemble of high-κ variants and of *wt* PNT4 is wider and bimodal (**Figures 6D** and **7D**). *wt* PNT4 displays two peaks centered at 23.13 Å and 54.09 Å (**Figure 7D**), while high-κ N_{TAIL} and high-κ PNT4 exhibit two peaks at 24.24

Å and 58.23 Å (N_{TAIL}) and 23.05 Å and 54.21 Å (PNT4). These data indicate that the scattering curves of these latter variants do not reflect a randomly distributed ensemble of conformations and R_g distributions thereby testifying their reproducibility (data not shown).

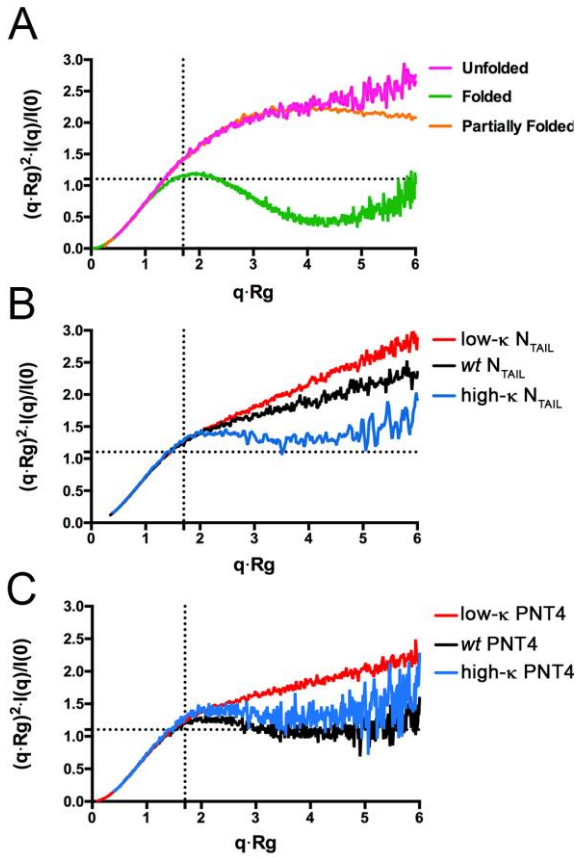


Figure 8. Dimensionless Kratky plots. A) Three proteins illustrative of three different conformations. Unfolded protein: TtASR1 (Hamdi et al., 2017); partially folded protein: Hendra virus V protein, (Salladini et al., 2017); folded protein: Hendra virus XD, (Erales et al., 2015). **B**) N_{TAIL} and **C**) PNT4 variants.

4. Discussion

The effects of charge distribution on the conformational properties of IDPs have been studied by comparing protein variants with identical amino acid composition but different linear distribution of charged residues. We have considered two sets of proteins, each of them including a natural protein (*wt*) and two synthetic variants (low-*κ* and high-*κ*) obtained by rational design.

The compaction trend accompanying the increase in *κ* value is consistently supported by different and independent techniques. The first line of support is represented by the higher proteolysis sensitivity exhibited by low-*κ* variants and, *viceversa*, the higher stability of high-*κ* variants. One can observe that different sensitivity to trypsin may result not only from conformational differences among protein variants, but also from different positions of basic residues along the sequence. This might occur for high-*κ* variants, where trypsin may digest fast the N-*terminus* that is enriched in Arg-Lys residues (Šlechtová et al., 2015), while leaving intact the C-*terminus* that is devoid of basic residues. On the contrary, we do not observe any rapid degradation of high-*κ* variants, suggesting that the observed differences in digestion patterns mainly reflect differences in conformational features of these proteins.

The experimental values of R_S, D_{max} and R_g for *wt* forms of both model proteins consistently show that charge clustering enhances protein compactness. The picture is further enriched by ESI-MS results, available at the moment for the lone NTAIL.

Compaction indexes (CI_{Rs}, CI_{Rg}) describe the changes of compactness independently of protein length. To take into account the changes of κ value, we have introduced an indicator of compaction responsiveness, namely ${}^{\kappa}CR_{Rs}$ or ${}^{\kappa}CR_{Rg}$ that refer respectively to the experimentally observed R_S and R_g and are obtained as the slope of linear regressions of CI on κ value (**Figure 9**). Due to the paucity of experimental points considered in each data set, ${}^{\kappa}CR_{Rs}$ can be regarded

as an oversimplification of the relationship between CI and κ. Nonetheless, we would like to underscore that it represents a first attempt to quantitively describe the experimental responsiveness of a protein to variations in the κ value.

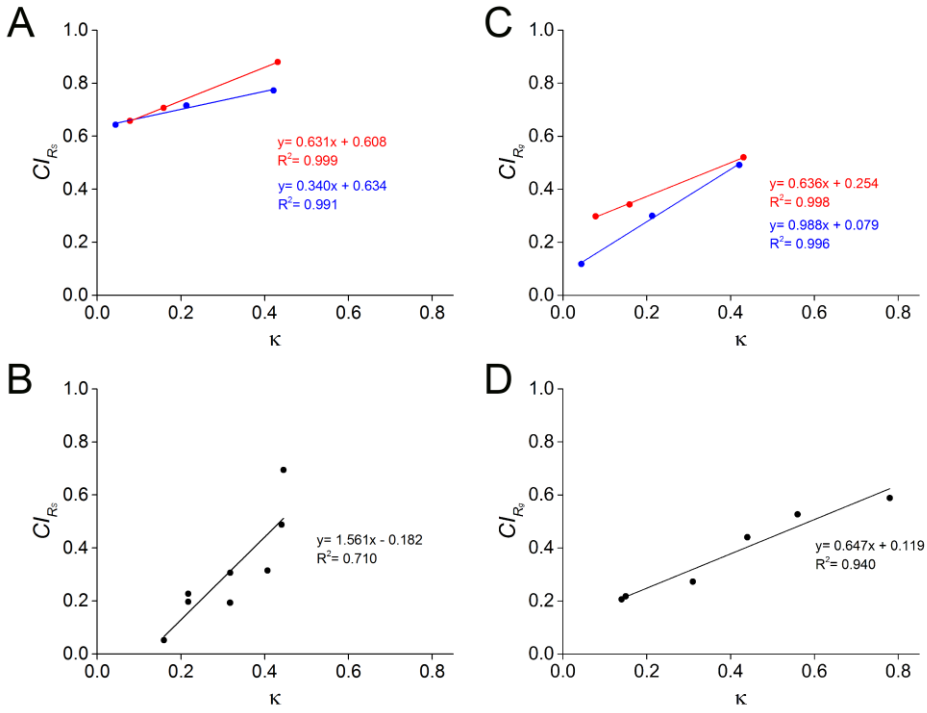


Figure 9. Linear regressions of CI_{RS} and CI_{Rg} on κ for N_{TAIL}, PNT4, $p27_{96-198}$ and NAM. Regression of CI_{RS} for N_{TAIL} (red), PNT4 (blue) (**A**) and NAM (**B**). Regression of CI_{Rg} for N_{TAIL} (red), PNT4 (blue) (**C**) and $p27_{96-198}$ (**D**). The equation of trend lines and R^2 are presented for each set of data. Data sets of N_{TAIL} and PNT4 were obtained from this work, data for NAM are from (Sherry et al., 2017), data for $p27_{96-198}$ data are from (Das et al., 2016).

$^{\kappa}CR_{Rs}$ and $^{\kappa}CR_{Rg}$ are very close to each other in the case of NTAIL. Conversely, and surprisingly, these two parameters are very dissimilar in the case of PNT4 (**Figure 9A** and **9C**, **Table 2**).

Specifically, PNT4 exhibits an exacerbated sensitivity to charge clustering in terms of Rg, whereas R_S appears resilient in responding to changes of κ values. Since the quality of the data is comparable for the two model proteins, and since SEC and SAXS are both reliable approaches, the peculiar behaviour of PNT4 can be regarded as reflecting inherent properties of this IDP.

Herein, we have also calculated and used the same parameter, $^{\kappa}CR$, to describe the responsiveness to κ changes of two previously investigated IDPs, namely $p27_{96-198}$ (Das et al., 2016) and NAM (Sherry et al., 2017). The values of $^{\kappa}CR$ for $p27_{96-198}$ and NAM are also presented in **Table 2** and **Figure 9B** and **9D**. Each protein seems to respond in a peculiar way to charge patterning. The regression of CI_{Rg} on κ value is linear in the case of p2796-198 (**Figure 9D**), and only roughly approximated to linearity in the case of $^{\kappa}CI_{Rg}$ for NAM (**Figure 9B**). Functional assays of permutated κ-variants of NAM have already highlighted a more cooperative response to charge clustering for this protein (Sherry et al., 2017). We reasoned that the differences in $^{\kappa}CR$ might depend on various sequence attributes, such as number of residues, hydrophobicity, proline content, NCPR and FCR. **Table 2** summarizes some of these sequence properties along with $^{\kappa}CR_{Rg}$ and $^{\kappa}CR_{Rg}$ values. All the considered proteins share similar hydrophobicity and sequence length, as well as very low |NCPR| values (< 0.05). The FCR and Pro percentage are the most fluctuating parameters. For instance, the NAM protein, which undergoes the most pronounced compaction, also exhibits the highest FCR and lowest Pro content. While it is evident that the compaction changes are not paralleled by variations in FCR, the Pro content appears as the most probable feature accounting for the responsiveness of conformational fluctuations (**Table 2**). Indeed, with the notable exception of PNT4 that is further discussed below, the

Pro content inversely correlates with $^{\kappa}$CR, irrespective of the analytical techniques used to obtain the experimental data. This interpretation agrees with the already postulated role of Pro frequency on the conformational properties and the responsiveness of IDPs to charge clustering (Das and Pappu, 2013; Das et al., 2015). Indeed, Pro is recognized as a disorder-promoting (Dunker et al., 2001a) and structure-breaker residue (Adzhubei and Sternberg, 1993), whose frequency well correlates with extended conformations (Marsh and Forman-Kay, 2010). These properties may be ascribed to the rigid and extended backbone dihedral angles it can adopt in the Ramachandran plot and that are not influenced by linear charge patterning. Hence a proline-rich protein exhibits an extended conformation more resilient to charges patterning than an equally sized and similarly disordered protein where Pro residues are less abundant. In this interpretative context, the Pro content of PNT4 (5.2 %) suggests that the conformational responsiveness of this protein is better described by its $^{\kappa}CR_{Rg}$ (0.988) than by its $^{\kappa}CR_{Rs}$ (0.340) that is much smaller than expected. The reasons underlying the unresponsiveness of the R_S of PNT4 to changes in κ value remain elusive so far and await future site-resolved studies that will unveil possible unique local features accounting for this peculiar behaviour.

The compaction of PNT4 seems entirely ascribable to the formation of tertiary contacts. For N_{TAIL}, the charge clustering not only leads to compaction but also to an increase in the content of secondary structure, as seen by far-UV CD spectroscopy. That the content in tertiary structure correlates with the content in regular secondary is in contrast with previous findings from us and others indicating that the latter is not a major determinant of protein compactness (Blocquel et al., 2012; Marsh and Forman-Kay, 2010).

It can be observed that N_{TAIL} contains a MoRE, partially conserved upon the protein re-design. Hence it is overall less disordered than PNT4, although its higher content of Pro.

These results therefore challenge the general assumption that the PMG state mainly arises from the occurrence of either long-range or short-range tertiary contacts rather than from local constraints imposed by transiently populated regular secondary structure elements and well illustrate the complexity of the conformational behaviour of IDPs.

Incidentally, Pondr-fit provides disorder profiles supporting that *wt* N_{TAIL} is less disordered than *wt* PNT4, but fails to detect the increased content in ordered structure, as judged from both the secondary structure content and the degree of compactness, of high-κ variants thus calling to further improvements of this predictor.

In conclusion, the results herein reported indicate that the distribution of opposite charges along the protein sequence affects the conformational properties of IDPs according to their overall composition and, in particular, to Pro content, which deserves a more systematic analysis. Coherent results were obtained with primary structures designed according to different constraints. In consideration of the distribution of κ values among natural IDRs (**Figure 10**) and of the high propensity to aggregate of high-κ variants, we also propose that sequence features within natural IDRs have evolved to ensure an optimal balance of sequence-encoded conformational properties, and prevention of aggregation. The evolutionary constraint may explain, together with entropic reasons (Das and Pappu, 2013), the preponderance among natural IDPs of κ values in the range of 0.1–0.4. The present findings not only shed light on the conformational behaviour of IDPs and on how this latter is encoded by the amino acid sequence, but are also expected to stimulate future studies aimed at rationally conceiving/designing non-natural IDPs with a desired degree of compactness.

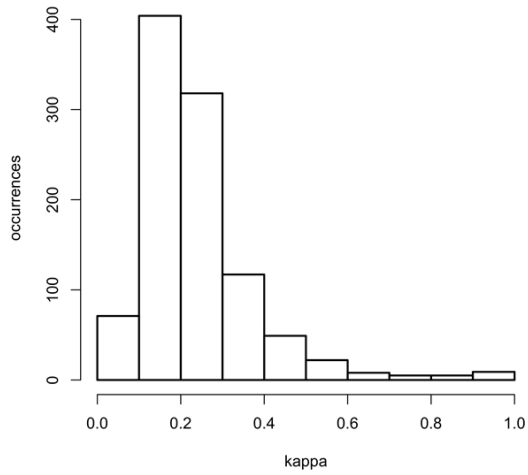


Figure 10. Frequency of κ within natural IDRs within the DisProt. All protein regions annotated in the DisProt 7.0 (Piovesan et al., 2017) were extracted, then redundant sequences and sequences of less than 20 residues were discarded. The κ value was computed using CIDER (in local mode) (Holehouse et al., 2017).

Acknowledgements

We acknowledge the European Synchrotron Radiation Facility for beamtime allocation and Bart van Laer and Petra Pernot for their assistance in using beamline BM29. We also thank Dr. Gerlind Sulzenbacher (AFMB lab) for efficiently managing the AFMB BAG. We thank Marco Mangiagalli for fruitful discussions and valuable suggestions, and Antoine Schramm (AFMB lab) for kindly having analyzed the distribution of κ value within DisProt entries and for having generated supplementary Figure S1. This work was partly supported by a grant Fondo di Ateneo of the University of Milano-Bicocca to SB. It was also partly supported by the CNRS (S.L.). G.T. acknowledges support by the Italian national PhD program. E.S. is supported by a joint doctoral fellowship from the Direction Générale de l'Armement (DGA) and Aix-Marseille University. C.S. benefited from an Assegno di Ricerca of the University of Milano-Bicocca.

Appendix

A1: From gene to disordered proteins

A1.1 Translation of IDPs

Interestingly, codon usage of disordered proteins differs from that used to encode globular proteins (Homma et al., 2016). Various hypothesis have been formulated to explain this observation. One hypothesis is related to translation efficiency. Free from structural constraints, IDPs and IDRs can accommodate more translational error than structured proteins and consequently codon usage is likely to be "less optimized" than in average genes (Pajkos et al., 2012). Translation efficiency hypothesis postulates that preferentially used codons are translated faster because the higher cellular concentrations of cognate tRNA, and *viceversa* (Ikemura, 1985). Recent data confutes translation efficiency hypothesis, indeed higher level of tRNA can inhibit RNA transcription (Agashe et al., 2012).

A1.2 Splicing of IDPs

Alternative splicing (AS) generates two or more protein isoforms from a single gene and it is a mechanism typical of multicellular organisms (Goren et al., 2006). The reason of the evolutionary success of alternative splicing can be found in the support to organism complexity and genome compactness. Indeed, alternative splicing allows to produce various forms of transcripts according to cell types, thus contributing to cell differentiation through well-controlled mechanisms, without increasing the proteome size (Schad et al., 2011) .Unfortunately, incorrect splicing of structured, multi-domain proteins can sometimes occur and it has usually strong negative structural effects on the translated proteins, often undergoing misfolding and aggregation (Demchenko, 2001). A systematic analysis of intron composition in a set of genes encoding for human proteins has led to the hypothesis that regions involved in AS are mainly encoding for IDRs (Romero et al., 2001). A more recent study also showed that splicing junctions

present a typical nucleotide composition which is likely translated as Gly, Ala, Pro and Arg, amino acids highly frequent in IDPs (Smithers et al., 2015).

A1.3 Evolution of structural intrinsic disorder

Different work demonstrate that proteome of multicellular eukaryotes is overall more "disordered" than the proteome of less complex organisms (Burra et al., 2010; Dunker et al., 2000; Feng et al., 2006; Ward et al., 2004).

The Earth formed about 4.5 billion years ago: some organic molecules could have been spontaneously produced from gases of the primitive reducing Earth atmosphere (Ferry and House, 2005). This starting hypothesis has been confirmed by a very famous experiment: various organic compounds, including some amino acids, were synthesize using non-organic compounds probably present in the early Earth atmosphere. Interestingly, not all the amino acids were synthesized, and this could mean that first proteins contained only few amino acids. This result should be read together with another interesting finding: the biosynthetic theory of the genetic code evolution suggests that the genetic code evolved from a simpler form that encodes few amino acids. One of the feature of genetic code is the redundancy and this is connected above all to the thirds nucleotide of the triplets. Starting from these observations, it has been proposed that the genetic code has evolved in two steps, the first implying the use of nucleotide "doublet", before the triplet code arose. Based on these and many other premises, it is possible to discriminate between "old" and "new" amino acids. In 2000, a Russian researcher combined 40 different single-factor criteria into a consensus scale and proposed the following temporal order of amino acids appearance in the genetic code: Gly/Ala, Val/Asn, Pro, Ser, Gln/Leu, Thr, Arg, Asp, Lys, Glu, Ile, Cys, His, Phe, Met, Tyr, Trp (Dunker et al., 2000). Even superficial analysis of this sequence reveals that many of the early amino acids are disordered-promoting, as they are very abundant in modern IDPs. On the other hand, the major order-promoting residues

were added to genetic code late. This strongly suggests that the primordial polypeptides were intrinsically disordered. This hypothesis is also confirmed considering that primitive Earth was characterised by very high temperature and IDP gene are riche of GC, so much resistant to high temperature (Yakovchuk et al., 2006). It is very unlikely that these disordered primordial polypeptides possessed catalytic activity (Poole et al., 1998). This hypothesis is in line with the "RNA-world theory" suggesting that during the evolution of enzymatic activity, catalysis was transferred from RNA first to ribonucleoprotein (RNP) and only then to protein. The global evolution of intrinsic disorder is characterized by a wavy pattern, where highly disordered primordial proteins with primarily RNA-chaperone activities were gradually substituted by the well- folded, highly ordered enzymes that evolved to catalyse the production of all the complex "goodies" crucial for the independent existence of the first cellular organism (Uversky, 2013a). Various studies have reported variegated scenarios to describe the evolutionary rates of ordered proteins and IDPs/IDRs in modern organism. In some cases, ordered and disordered domains in the same protein were shown to possess similar degree of conservation and co-evolution (Chemes et al., 2012) Several studies suggest that IDPs/IDRs present high evolutionary rates while maintaining the structural disorder and their physiological functions (Brown et al., 2009; Brown et al., 2002; Chen et al., 2011; Lin et al., 2007). Since the degree of positive Darwinian selection appears significantly higher in IDP/IDRs than in structured proteins, it was hypothesized that structural disorder may be required to produce genetic variations.

Related to IDPs evolution, recently some group analysed IDPs at DNA level.

Disordered proteins/regions are rich of Gly, Ala, Pro and Arg (Uversky, 2011), often encoded by triplets rich of GC. There is a strong correlation between the prokaryotic optimal growth at higher temperatures and the GC content in their whole genome (Musto et al., 2004; Musto et al., 2006). A differential frequency

of GC has been observed also in many eukaryotes genomes, where it correlates with altered mutation and recombination frequency (Costantini et al., 2013). These observations have stimulated several computational studies aimed at finding a relationship between GC content, the genome size, localization of GC-rich regions and the presence of gene encoding for IDPs (Galea et al., 2006; Peng et al., 2016; Schad et al., 2011). Among 296 prokaryote genomes analysed, the highest disorder protein disorder level is associated to highest GC content, whereas in archaea the highest disorder content is also associated to a small genome size, thus the highest number of disordered proteins or regions has been found in small, GC-rich genomes (Peng et al., 2016).

In conclusion, tandem repeats of GC are enriched and they are more frequent in genes coding for IDPs. The genetic instability of repetitive genomic regions, in combination with the structurally permissive nature of IDRs, might have driven the increase of the amount of disorder during the evolution, as “orphan proteins” also demonstrate. Indeed, genes of orphan proteins show high level of GC content (Basile et al., 2017).

5. References

Adzhubei, A.A., and Sternberg, M.J.E. (1993). Left-handed polyproline II helices commonly occur in globular proteins. Journal of molecular biology *229*, 472-493.

Agashe, D., Martinez-Gomez, N.C., Drummond, D.A., and Marx, C.J. (2012). Good codons, bad transcript: large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. Molecular biology and evolution *30*, 549-560.

Agostini, F., Vendruscolo, M., and Tartaglia, G.G. (2012). Sequence-based prediction of protein solubility. Journal of molecular biology *421*, 237-241.

Arrondo, J.L.R., and Goñi, F.M. (1999). Structure and dynamics of membrane proteins as studied by infrared spectroscopy. Progress in biophysics and molecular biology *72*, 367-405.

Arrondo, J.L.R., Muga, A., Castresana, J., and Goñi, F.M. (1993). Quantitative studies of the structure of proteins in solution by Fourier-transform infrared spectroscopy. Progress in biophysics and molecular biology *59*, 23-56.

Atkins, P., and De Paula, J. (2013). Elements of physical chemistry (Oxford University Press, USA).

Babu, M.M., Kriwacki, R.W., and Pappu, R.V. (2012). Versatility from protein disorder. Science *337*, 1460-1461.

Balázs, A., Csizmok, V., Buday, L., Rakács, M., Kiss, R., Bokor, M., Udupa, R., Tompa, K., and Tompa, P. (2009). High levels of structural disorder in scaffold proteins as exemplified by a novel neuronal protein, CASK-interactive protein1. The FEBS journal *276*, 3744-3756.

Barth, A. (2007). Infrared spectroscopy of proteins. Biochimica et Biophysica Acta (BBA)-Bioenergetics *1767*, 1073-1101.

Basile, W., Sachenkova, O., Light, S., and Elofsson, A. (2017). High GC content causes orphan proteins to be intrinsically disordered. PLoS computational biology *13*, e1005375.

Bath, A., and Zscherp, C. (2002). What vibrations tell about proteins. Q. Rev. Biophys *35*, 369-430.

Blocquel, D., Habchi, J., Gruet, A., Blangy, S., and Longhi, S. (2012). Compaction and binding properties of the intrinsically disordered C-terminal domain of Henipavirus nucleoprotein as unveiled by deletion studies. Molecular BioSystems *8*, 392-410.

Boehr, D.D., Nussinov, R., and Wright, P.E. (2009). The role of dynamic conformational ensembles in biomolecular recognition. Nature chemical biology *5*, 789-796.

Bolen, D.W., and Rose, G.D. (2008). Structure and energetics of the hydrogen-bonded backbone in protein folding. Annu. Rev. Biochem. *77*, 339-362.

Bourhis, J.-M., Johansson, K., Receveur-Bréchot, V., Oldfield, C.J., Dunker, K.A., Canard, B., and Longhi, S. (2004). The C-terminal domain of measles virus nucleoprotein belongs to the class of intrinsically disordered proteins that fold upon binding to their physiological partner. Virus research *99*, 157-167.

Brangwynne, C.P., Tompa, P., and Pappu, R.V. (2015). Polymer physics of intracellular phase transitions. Nature Physics *11*, 899-904.

Breydo, L., Redington, J.M., and Uversky, V.N. (2017). Effects of Intrinsic and Extrinsic Factors on Aggregation of Physiologically Important Intrinsically Disordered Proteins. In International review of cell and molecular biology (Elsevier), pp. 145-185.

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

Brown, C.J., Johnson, A.K., and Daughdrill, G.W. (2009). Comparing models of evolution for ordered and disordered proteins. Molecular biology and evolution *27*, 609-621.

Brown, C.J., Takayama, S., Campen, A.M., Vise, P., Marshall, T.W., Oldfield, C.J., Williams, C.J., and Keith Dunker, A. (2002). Evolutionary rate heterogeneity in proteins with long disordered regions. Journal of molecular evolution *55*, 104-110.

Burra, P.V., Kalmar, L., and Tompa, P. (2010). Reduction in structural disorder and functional complexity in the thermal adaptation of prokaryotes. PloS one *5*, e12069.

Campen, A., Williams, R.M., Brown, C.J., Meng, J., Uversky, V.N., and Dunker, A.K. (2008). TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. Protein and peptide letters *15*, 956-963.

Campos, F., Guillén, G., Reyes, J.L., and Covarrubias, A.A. (2011). A general method of protein purification for recombinant unstructured non-acidic proteins. Protein expression and purification *80*, 47-51.

Chan, P., Curtis, R.A., and Warwicker, J. (2013). Soluble expression of proteins correlates with a lack of positively-charged surface. Scientific reports *3*, 3333.

Charneski, C.A., and Hurst, L.D. (2013). Positively charged residues are the major determinants of ribosomal velocity. PLoS Biol *11*, e1001508.

Chemes, L.B., Glavina, J., Alonso, L.G., Marino-Buslje, C., de Prat-Gay, G., and Sánchez, I.E. (2012). Sequence evolution of the intrinsically disordered and globular domains of a model viral oncoprotein. PLoS One *7*, e47661.

Chen, J.W., Romero, P., Uversky, V.N., and Dunker, A.K. (2006). Conservation of intrinsic disorder in protein domains and families: II. functions of conserved disorder. Journal of proteome research *5*, 888-898.

Chen, S.C.-C., Chuang, T.-J., and Li, W.-H. (2011). The relationships among microRNA regulation, intrinsically disordered regions, and other indicators of protein evolutionary rate. Molecular biology and evolution *28*, 2513-2520.

Chiti, F., Calamai, M., Taddei, N., Stefani, M., Ramponi, G., and Dobson, C.M. (2002). Studies of the aggregation of mutant proteins in vitro provide insights into

the genetics of amyloid diseases. Proceedings of the National Academy of Sciences *99*, 16419-16426.

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

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

Costa, S., Almeida, A., Castro, A., and Domingues, L. (2014). Fusion tags for protein solubility, purification, and immunogenicity in Escherichia coli: the novel Fh8 system. Recombinant protein expression in microbial systems, 24.

Costantini, S., Sharma, A., Raucci, R., Costantini, M., Autiero, I., and Colonna, G. (2013). Genealogy of an ancient protein family: the Sirtuins, a family of disordered members. BMC evolutionary biology *13*, 60.

Csizmók, V., Szőllősi, E., Friedrich, P., and Tompa, P. (2006). A novel two-dimensional electrophoresis technique for the identification of intrinsically unstructured proteins. Molecular & Cellular Proteomics *5*, 265-273.

Das, R.K., Huang, Y., Phillips, A.H., Kriwacki, R.W., and Pappu, R.V. (2016). Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling. Proceedings of the National Academy of Sciences *113*, 5616-5621.

Das, R.K., and Pappu, R.V. (2013). Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proceedings of the National Academy of Sciences *110*, 13392-13397.

Das, R.K., Ruff, K.M., and Pappu, R.V. (2015). Relating sequence encoded information to form and function of intrinsically disordered proteins. Current opinion in structural biology *32*, 102-112.

Davey, N.E., Van Roey, K., Weatheritt, R.J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H., and Gibson, T.J. (2012). Attributes of short linear motifs. Molecular BioSystems *8*, 268-281.

De Baets, G., Van Durme, J., van der Kant, R., Schymkowitz, J., and Rousseau, F. (2015). Solubis: optimize your protein. Bioinformatics *31*, 2580-2582.

Demchenko, A.P. (2001). Recognition between flexible protein molecules: induced and assisted folding. Journal of molecular recognition *14*, 42-61.

Diella, F., Haslam, N., Chica, C., Budd, A., Michael, S., Brown, N.P., Travé, G., and Gibson, T.J. (2008). Understanding eukaryotic linear motifs and their role in cell signaling and regulation. Front Biosci *13*, 603.

Dosztányi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics *21*, 3433-3434.

Dunker, A.K., Babu, M.M., Barbar, E., Blackledge, M., Bondos, S.E., Dosztányi, Z., Dyson, H.J., Forman-Kay, J., Fuxreiter, M., and Gsponer, J. (2013). What's in a name? Why these proteins are intrinsically disordered: Why these proteins are intrinsically disordered. Intrinsically Disordered Proteins *1*, e24157.

Dunker, A.K., Brown, C.J., and Obradovic, Z. (2002). Identification and functions of usefully disordered proteins. Advances in protein chemistry *62*, 25-49.

Dunker, A.K., Cortese, M.S., Romero, P., Iakoucheva, L.M., and Uversky, V.N. (2005a). Flexible nets. Febs Journal *272*, 5129-5148.

Dunker, A.K., Cortese, M.S., Romero, P., Iakoucheva, L.M., and Uversky, V.N. (2005b). Flexible nets. The FEBS journal *272*, 5129-5148.

Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., and Hipps, K.W. (2001a). Intrinsically disordered protein. Journal of Molecular Graphics and Modelling *19*, 26-59.

Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.R., Hipps, K.W., *et al.* (2001b). Intrinsically disordered protein. Journal of Molecular Graphics & Modelling *19*, 26-59.

Dunker, A.K., and Obradovic, Z. (2001). The protein trinity—linking function and disorder. Nature biotechnology *19*, 805-806.

Dunker, A.K., Oldfield, C.J., Meng, J., Romero, P., Yang, J.Y., Chen, J.W., Vacic, V., Obradovic, Z., and Uversky, V.N. (2008). The unfoldomics decade: an update on intrinsically disordered proteins. BMC genomics *9*, S1.

Dunker, A.K., Romero, P., Obradovic, Z., Garner, E.C., and Brown, C.J. (2000). Intrinsic protein disorder in complete genomes. Genome Informatics *11*, 161-171.

Dyson, H.J., and Wright, P.E. (2005). Intrinsically unstructured proteins and their functions. Nature reviews. Molecular cell biology *6*, 197.

Edbauer, D., and Haass, C. (2016). An amyloid-like cascade hypothesis for C9orf72 ALS/FTD. Current opinion in neurobiology *36*, 99-106.

Erales, J., Beltrandi, M., Roche, J., Maté, M., and Longhi, S. (2015). Insights into the Hendra virus N TAIL–XD complex: evidence for a parallel organization of the helical MoRE at the XD surface stabilized by a combination of hydrophobic and polar interactions. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics *1854*, 1038-1053.

Felitsky, D.J., Lietzow, M.A., Dyson, H.J., and Wright, P.E. (2008). Modeling transient collapsed states of an unfolded protein to provide insights into early folding events. Proceedings of the National Academy of Sciences *105*, 6278-6283.

Feng, Z.-P., Zhang, X., Han, P., Arora, N., Anders, R.F., and Norton, R.S. (2006). Abundance of intrinsically unstructured proteins in P. falciparum and other apicomplexan parasite proteomes. Molecular and biochemical parasitology *150*, 256-267.

Ferry, J.G., and House, C.H. (2005). The stepwise evolution of early life driven by energy conservation. Molecular biology and evolution *23*, 1286-1292.

Flock, T., Weatheritt, R.J., Latysheva, N.S., and Babu, M.M. (2014). Controlling entropy to tune the functions of intrinsically disordered regions. Current opinion in structural biology *26*, 62-72.

Flory, P., and Volkenstein, M. (1969). Statistical mechanics of chain molecules. (Wiley Online Library).

Forman-Kay, J.D., and Mittag, T. (2013). From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. Structure *21*, 1492-1499.

Fuxreiter, M. (2012). Fuzziness: linking regulation to protein dynamics. Molecular BioSystems *8*, 168-177.

Fuxreiter, M., Tompa, P., and Simon, I. (2007). Local structural disorder imparts plasticity on linear motifs. Bioinformatics *23*, 950-956.

Galea, C.A., Pagala, V.R., Obenauer, J.C., Park, C.-G., Slaughter, C.A., and Kriwacki, R.W. (2006). Proteomic studies of the intrinsically unstructured mammalian proteome. Journal of proteome research *5*, 2839-2848.

Gast, K., Damaschun, H., Eckert, K., Schulze-Forster, K., Maurer, H.R., Mueller-Frohne, M., Zirwer, D., Czarnecki, J., and Damaschun, G. (1995). Prothymosin. alpha.: A Biologically Active Protein with Random Coil Conformation. Biochemistry *34*, 13211-13218.

Gast, K., Damaschun, H., Misselwitz, R., Müller-Frohne, M., Zirwer, D., and Damaschun, G. (1994). Compactness of protein molten globules: temperature-induced structural changes of the apomyoglobin folding intermediate. European biophysics journal *23*, 297-305.

Glatter, O. (1977). Data evaluation in small angle scattering: calculation of the radial electron density distribution by means of indirect Fourier transformation. Acta Physica Austriaca *47*, 83-102.

Glatter, O., and Kratky, O. (1982). Small angle X-ray scattering (Academic press).

Goormaghtigh, E., Cabiaux, V., and Ruysschaert, J.-M. (1994). Determination of soluble and membrane protein structure by Fourier transform infrared spectroscopy. In Physicochemical methods in the study of biomembranes (Springer), pp. 405-450.

Goormaghtigh, E., Raussens, V., and Ruysschaert, J.-M. (1999). Attenuated total reflection infrared spectroscopy of proteins and lipids in biological membranes. Biochimica et Biophysica Acta (BBA)-Reviews on Biomembranes *1422*, 105-185.

Goren, A., Ram, O., Amit, M., Keren, H., Lev-Maor, G., Vig, I., Pupko, T., and Ast, G. (2006). Comparative analysis identifies exonic splicing regulatory sequences—the complex definition of enhancers and silencers. Molecular cell *22*, 769-781.

Habchi, J., and Longhi, S. (2012). Structural disorder within paramyxovirus nucleoproteins and phosphoproteins. Molecular Biosystems *8*, 69-81.

Habchi, J., Mamelli, L., Darbon, H., and Longhi, S. (2010). Structural disorder within Henipavirus nucleoprotein and phosphoprotein: from predictions to experimental assessment. PLoS One *5*, e11684.

Habchi, J., Tompa, P., Longhi, S., and Uversky, V.N. (2014). Introducing protein intrinsic disorder. Chemical reviews *114*, 6561-6588.

Halfmann, R. (2016). A glass menagerie of low complexity sequences. Current opinion in structural biology *38*, 18-25.

Hamdi, K., Salladini, E., O'Brien, D.P., Brier, S., Chenal, A., Yacoubi, I., and Longhi, S. (2017). Structural disorder and induced folding within two cereal, ABA stress and ripening (ASR) proteins. Scientific Reports *7*, 15544.

Hiller, S., Wider, G., Imbach, L.L., and Wüthrich, K. (2008). Interactions with Hydrophobic Clusters in the Urea-Unfolded Membrane Protein OmpX. Angewandte Chemie International Edition *47*, 977-981.

Hipp, N.J., Groves, M.L., Custer, J.H., and McMeekin, T.L. (1952). Separation of α-, β-and γ-Casein1. Journal of Dairy Science *35*, 272-281.

Hofmann, H., Golbik, R.P., Ott, M., Hübner, C.G., and Ulbrich-Hofmann, R. (2008). Coulomb forces control the density of the collapsed unfolded state of barstar. Journal of molecular biology *376*, 597-605.

Holehouse, A.S., Das, R.K., Ahad, J.N., Richardson, M.O.G., and Pappu, R.V. (2017). CIDER: resources to analyze sequence-ensemble relationships of intrinsically disordered proteins. Biophysical journal *112*, 16-21.

Holehouse, A.S., and Pappu, R.V. (2018). Functional Implications of Intracellular Phase Transitions. Biochemistry.

Homma, K., Noguchi, T., and Fukuchi, S. (2016). Codon usage is less optimized in eukaryotic gene segments encoding intrinsically disordered regions than in those encoding structural domains. Nucleic acids research *44*, 10051-10061.

Højgaard, C., Kofoed, C., Espersen, R., Johansson, K.E., Villa, M., Willemoës, M., Lindorff-Larsen, K., Teilum, K., and Winther, J.R. (2016). A soluble, folded protein without charged amino acid residues. Biochemistry *55*, 3949-3956.

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. Molecular biology and evolution *2*, 13-34.

Isom, D.G., Castañeda, C.A., and Cannon, B.R. (2011). Large shifts in pKa values of lysine residues buried inside a protein. Proceedings of the National Academy of Sciences *108*, 5260-5265.

Jeffery, C.J. (2003). Multifunctional proteins: examples of gene sharing. Annals of medicine *35*, 28-35.

Jeffery, C.J. (2004). Molecular mechanisms for multitasking: recent crystal structures of moonlighting proteins. Current opinion in structural biology *14*, 663-668.

Kar, G., Keskin, O., Gursoy, A., and Nussinov, R. (2010). Allostery and population shift in drug discovery. Current opinion in pharmacology *10*, 715-722.

Karlin, D., Longhi, S., Receveur, V., and Canard, B. (2002). The N-terminal domain of the phosphoprotein of morbilliviruses belongs to the natively unfolded class of proteins. Virology *296*, 251-262.

Kim, S.-H., Shin, D.H., Choi, I.-G., Schulze-Gahmen, U., Chen, S., and Kim, R. (2003). Structure-based functional inference in structural genomics. Journal of structural and functional genomics *4*, 129-135.

Kim, W., and Hecht, M.H. (2006). Generic hydrophobic residues are sufficient to promote aggregation of the Alzheimer's Aβ42 peptide. Proceedings of the National Academy of Sciences *103*, 15824-15829.

Kim, Y.-I., Burton, R.E., Burton, B.M., Sauer, R.T., and Baker, T.A. (2000). Dynamics of substrate denaturation and translocation by the ClpXP degradation machine. Molecular cell *5*, 639-648.

Konarev, P.V., Volkov, V.V., Sokolova, A.V., Koch, M.H.J., and Svergun, D.I. (2003). PRIMUS: a Windows PC-based system for small-angle scattering data analysis. Journal of applied crystallography *36*, 1277-1282.

Korshavn, K.J., Satriano, C., Lin, Y., Zhang, R., Dulchavsky, M., Bhunia, A., Ivanova, M.I., Lee, Y.-H., La Rosa, C., and Lim, M.H. (2017). JBC Papers in Press. Published on February 1, 2017 as Manuscript M116. 764092.

Kozlowski, L.P. (2016). IPC–Isoelectric Point Calculator. Biology direct *11*, 55.

Kramer, R.M., Shende, V.R., Motl, N., Pace, C.N., and Scholtz, J.M. (2012). Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility. Biophysical journal *102*, 1907-1915.

Laemmli, U.K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. nature *227*, 680-685.

Lawrence, M.S., Phillips, K.J., and Liu, D.R. (2007). Supercharging proteins can impart unusual resilience. Journal of the American Chemical Society *129*, 10110-10112.

Li, H., Helling, R., Tang, C., and Wingreen, N. (1996). Emergence of preferred structures in a simple model of protein folding. Science *273*, 666.

Lin, Y.-S., Hsu, W.-L., Hwang, J.-K., and Li, W.-H. (2007). Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. Molecular biology and evolution *24*, 1005-1011.

Lindberg, D.J., Wesén, E., Björkeroth, J., Rocha, S., and Esbjörner, E.K. (2017). Lipid membranes catalyse the fibril formation of the amyloid-β (1–42) peptide through lipid-fibril interactions that reinforce secondary pathways. Biochimica et Biophysica Acta (BBA)-Biomembranes *1859*, 1921-1929.

Linding, R., Schymkowitz, J., Rousseau, F., Diella, F., and Serrano, L. (2004). A comparative study of the relationship between protein structure and β-aggregation in globular and intrinsically disordered proteins. Journal of molecular biology *342*, 345-353.

Lise, S., and Jones, D.T. (2005). Sequence patterns associated with disordered regions in proteins. PROTEINS: Structure, Function, and Bioinformatics *58*, 144-150.

Liu, C.L., Tangsombatvisit, S., Rosenberg, J.M., Mandelbaum, G., Gillespie, E.C., Gozani, O.P., Alizadeh, A.A., and Utz, P.J. (2012). Specific post-translational histone modifications of neutrophil extracellular traps as immunogens and potential targets of lupus autoantibodies. Arthritis research & therapy *14*, R25.

Liu, J., Perumal, N.B., Oldfield, C.J., Su, E.W., Uversky, V.N., and Dunker, A.K. (2006). Intrinsic disorder in transcription factors. Biochemistry *45*, 6873-6888.

Livernois, A.M., Hnatchuk, D.J., Findlater, E.E., and Graether, S.P. (2009). Obtaining highly purified intrinsically disordered protein by boiling lysis and single step ion exchange. Analytical biochemistry *392*, 70-76.

Loeb, J. (1918). Amphoteric colloids ii. Volumetric analysis of ion-protein compounds; the significance of the isoelectric point for the purification of amphoteric colloids. The Journal of general physiology *1*, 237-254.

Longhi, S., Bloyet, L.-M., Gianni, S., and Gerlier, D. (2017). How order and disorder within paramyxoviral nucleoproteins and phosphoproteins orchestrate the molecular interplay of transcription and replication. Cellular and Molecular Life Sciences, 1-28.

Longhi, S., Receveur-Bréchot, V., Karlin, D., Johansson, K., Darbon, H., Bhella, D., Yeo, R., Finet, S., and Canard, B. (2003). The C-terminal domain of the measles virus nucleoprotein is intrinsically disordered and folds upon binding to the C-terminal moiety of the phosphoprotein. Journal of Biological Chemistry *278*, 18638-18648.

Ma, B., and Nussinov, R. (2009). Regulating highly dynamic unstructured proteins and their coding mRNAs. Genome biology *10*, 204.

Mandell, D.J., Chorny, I., Groban, E.S., Wong, S.E., Levine, E., Rapp, C.S., and Jacobson, M.P. (2007). Strengths of hydrogen bonds involving phosphorylated amino acid side chains. Journal of the American Chemical Society *129*, 820-827.

Mao, A.H., Crick, S.L., Vitalis, A., Chicoine, C.L., and Pappu, R.V. (2010). Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. Proceedings of the National Academy of Sciences *107*, 8183-8188.

Marsh, J.A., and Forman-Kay, J.D. (2010). Sequence determinants of compaction in intrinsically disordered proteins. Biophysical journal *98*, 2383-2390.

Mohan, A., Oldfield, C.J., Radivojac, P., Vacic, V., Cortese, M.S., Dunker, A.K., and Uversky, V.N. (2006). Analysis of molecular recognition features (MoRFs). Journal of molecular biology *362*, 1043-1059.

Molliex, A., Temirov, J., Lee, J., Coughlin, M., Kanagaraj, A.P., Kim, H.J., Mittag, T., and Taylor, J.P. (2015). Phase separation by low complexity domains

promotes stress granule assembly and drives pathological fibrillization. Cell *163*, 123-133.

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

Murray, I.V.J., Liu, L., Komatsu, H., Uryu, K., Xiao, G., Lawson, J.A., and Axelsen, P.H. (2007). Membrane-mediated amyloidogenesis and the promotion of oxidative lipid damage by amyloid β proteins. Journal of biological chemistry *282*, 9335-9345.

Musto, H., Naya, H., Zavala, A., Romero, H., Alvarez-Valín, F., and Bernardi, G. (2004). Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. FEBS letters *573*, 73-77.

Musto, H., Naya, H., Zavala, A., Romero, H., Alvarez-Valín, F., and Bernardi, G. (2006). Genomic GC level, optimal growth temperature, and genome size in prokaryotes. Biochemical and biophysical research communications *347*, 1-3.

Möglich, A., Joder, K., and Kiefhaber, T. (2006). End-to-end distance distributions and intrachain diffusion constants in unfolded polypeptide chains indicate intramolecular hydrogen bond formation. Proceedings of the National Academy of Sciences *103*, 12394-12399.

Müller-Späth, S., Soranno, A., Hirschfeld, V., Hofmann, H., Rüegger, S., Reymond, L., Nettels, D., and Schuler, B. (2010). Charge interactions can dominate the dimensions of intrinsically disordered proteins. Proceedings of the National Academy of Sciences *107*, 14609-14614.

Natalello, A., Ami, D., and Doglia, S.M. (2012). Fourier transform infrared spectroscopy of intrinsically disordered proteins: measurement procedures and data analyses. Intrinsically Disordered Protein Analysis: Volume 1, Methods and Experimental Tools, 229-244.

Natalello, A., and Doglia, S.M. (2010). Intrinsically disordered proteins and induced folding studied by fourier transform infrared spectroscopy. Instrumental analysis of intrinsically disordered proteins, VN Uversky and S. Longhi, Editors. 2010, Wiley, 225-252.

Natalello, A., Relini, A., Penco, A., Halabelian, L., Bolognesi, M., Doglia, S.M., and Ricagno, S. (2015). Wild type beta-2 microglobulin and DE loop mutants display a common fibrillar architecture. PloS one *10*, e0122449.

Natalello, A., Santambrogio, C., and Grandori, R. (2017). Are Charge-State Distributions a Reliable Tool Describing Molecular Ensembles of Intrinsically Disordered Proteins by Native MS? Journal of The American Society for Mass Spectrometry *28*, 21-28.

Neduva, V., Linding, R., Su-Angrand, I., Stark, A., De Masi, F., Gibson, T.J., Lewis, J., Serrano, L., and Russell, R.B. (2005). Systematic discovery of new recognition peptides mediating protein interaction networks. PLoS biology *3*, e405.

Niwa, T., Ying, B.-W., Saito, K., Jin, W., Takada, S., Ueda, T., and Taguchi, H. (2009). Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. Proceedings of the National Academy of Sciences *106*, 4201-4206.

Nordén, B. (1997). Circular dichroism and linear dichroism, Vol 1 (Oxford University Press, USA).

Pajkos, M., Mészáros, B., Simon, I., and Dosztányi, Z. (2012). Is there a biological cost of protein disorder? Analysis of cancer-associated mutations. Molecular BioSystems *8*, 296-307.

Parravicini, F., Brocca, S., and Lotti, M. (2015). Evaluation of the Conformational Stability of Recombinant Desulfurizing Enzymes from a Newly Isolated Rhodococcus sp. Molecular biotechnology, 1-11.

Parravicini, F., Natalello, A., Papaleo, E., De Gioia, L., Doglia, S.M., Lotti, M., and Brocca, S. (2013). Reciprocal Influence of Protein Domains in the Cold-Adapted Acyl Aminoacyl Peptidase from Sporosarcina psychrophila. Plos One *8*, 11.

Pauwels, K., Lebrun, P., and Tompa, P. (2017). To be disordered or not to be disordered: is that still a question for proteins in the cell? Cellular and Molecular Life Sciences, 1-20.

Peng, K., Radivojac, P., Vucetic, S., Dunker, A.K., and Obradovic, Z. (2006). Length-dependent prediction of protein intrinsic disorder. BMC bioinformatics *7*, 208.

Peng, Z., Uversky, V.N., and Kurgan, L. (2016). Genes encoding intrinsic disorder in Eukaryota have high GC content. Intrinsically Disordered Proteins *4*, e1262225.

Petoukhov, M.V., Franke, D., Shkumatov, A.V., Tria, G., Kikhney, A.G., Gajda, M., Gorba, C., Mertens, H.D.T., Konarev, P.V., and Svergun, D.I. (2012). New developments in the ATSAS program package for small-angle scattering data analysis. Journal of applied crystallography *45*, 342-350.

Piovesan, D., Tabaro, F., Mičetić, I., Necci, M., Quaglia, F., Oldfield, C.J., Aspromonte, M.C., Davey, N.E., Davidović, R., and Dosztányi, Z. (2017). DisProt 7.0: a major update of the database of disordered proteins. Nucleic acids research *45*, D219-D227.

Poole, A.M., Jeffares, D.C., and Penny, D. (1998). The path from the RNA world. Journal of Molecular Evolution *46*, 1-17.

Prabakaran, R., Goel, D., Kumar, S., and Gromiha, M.M. (2017). Aggregation prone regions in human proteome: Insights from large-scale data analyses. Proteins: Structure, Function, and Bioinformatics *85*, 1099-1118.

Prilusky, J., Felder, C.E., Zeev-Ben-Mordehai, T., Rydberg, E.H., Man, O., Beckmann, J.S., Silman, I., and Sussman, J.L. (2005). FoldIndex©: a simple tool

to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics *21*, 3435-3438.

Rambaran, R.N., and Serpell, L.C. (2008). Amyloid fibrils. Prion.

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

Reynolds, J.A., and Tanford, C. (1970). Binding of dodecyl sulfate to proteins at high binding ratios. Possible implications for the state of proteins in biological membranes. Proceedings of the National Academy of Sciences *66*, 1002-1007.

Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., Garner, E., Guilliot, S., and Dunker, A.K. Thousands of proteins likely to have long disordered regions. pp. 437-448.

Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J., and Dunker, A.K. (2001). Sequence complexity of disordered protein. Proteins: Structure, Function, and Bioinformatics *42*, 38-48.

Salladini, E., Delauzun, V., and Longhi, S. (2017). The Henipavirus V protein is a prevalently unfolded protein with a zinc-finger domain involved in binding to DDB1. Molecular BioSystems.

Sambi, I., Gatti-Lafranconi, P., Longhi, S., and Lotti, M. (2010). How disorder influences order and vice versa - mutual effects in fusion proteins containing an intrinsically disordered and a globular protein. Febs Journal *277*, 4438-4451.

Sant'Anna, R., Braga, C., Varejão, N., Pimenta, K.M., Graña-Montes, R., Alves, A., Cortines, J., Cordeiro, Y., Ventura, S., and Foguel, D. (2014). The importance of a gatekeeper residue on the aggregation of transthyretin implications for transthyretin-related amyloidoses. Journal of Biological Chemistry *289*, 28324-28337.

Santner, A.A., Croy, C.H., Vasanwala, F.H., Uversky, V.N., Van, Y.-Y.J., and Dunker, A.K. (2012). Sweeping away protein aggregation with entropic bristles:

intrinsically disordered protein fusions enhance soluble expression. Biochemistry *51*, 7250-7262.

Schad, E., Tompa, P., and Hegyi, H. (2011). The relationship between proteome size, structural disorder and organism complexity. Genome biology *12*, R120.

Schmittschmitt, J.P., and Scholtz, J.M. (2003). The role of protein stability, solubility, and net charge in amyloid fibril formation. Protein Science *12*, 2374-2378.

Sherry, K.P., Das, R.K., Pappu, R.V., and Barrick, D. (2017). Control of transcriptional activity by design of charge patterning in the intrinsically disordered RAM region of the Notch receptor. Proceedings of the National Academy of Sciences *114*, E9243-E9252.

Shoemaker, K.R., Kim, P.S., York, E.J., Stewart, J.M., and Baldwin, R.L. (1987). Tests of the helix dipole model for stabilization of α-helices.

Sickmeier, M., Hamilton, J.A., LeGall, T., Vacic, V., Cortese, M.S., Tantos, A., Szabo, B., Tompa, P., Chen, J., and Uversky, V.N. (2006). DisProt: the database of disordered proteins. Nucleic acids research *35*, D786-D793.

Simon, J.R., Carroll, N.J., Rubinstein, M., Chilkoti, A., and López, G.P. (2017a). Programming molecular self-assembly of intrinsically disordered proteins containing sequences of low complexity. Nature Chemistry.

Simon, J.R., Carroll, N.J., Rubinstein, M., Chilkoti, A., and López, G.P. (2017b). Programming molecular self-assembly of intrinsically disordered proteins containing sequences of low complexity. Nature Chemistry *9*, 509-515.

Smithers, B., Oates, M.E., and Gough, J. (2015). Splice junctions are constrained by protein disorder. Nucleic acids research *43*, 4814-4822.

Srinivasan, N., Bhagawati, M., Ananthanarayanan, B., and Kumar, S. (2014). Stimuli-sensitive intrinsically disordered protein brushes. Nature communications *5*, 5145.

Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T.A., and Soreq, H. (2005). Function of alternative splicing. Gene *344*, 1-20.

Studier, F.W. (2005). Protein production by auto-induction in high-density shaking cultures. Protein Expression and Purification *41*, 207-234.

Su, Y., Zou, Z., Feng, S., Zhou, P., and Cao, L. (2007). The acidity of protein fusion partners predominantly determines the efficacy to improve the solubility of the target proteins expressed in Escherichia coli. Journal of biotechnology *129*, 373-382.

Susi, H., and Byler, D.M. (1986). [13] Resolution-enhanced fourier transform infrared spectroscopy of enzymes. Methods in enzymology *130*, 290-311.

Svergun, D.I. (1992). Determination of the regularization parameter in indirect-transform methods using perceptual criteria. Journal of applied crystallography *25*, 495-503.

Šlechtová, T., Gilar, M., Kalíková, K.t., and Tesařová, E. (2015). Insight into trypsin miscleavage: comparison of kinetic constants of problematic peptide sequences. Analytical chemistry *87*, 7636-7643.

Tamm, L.K., and Tatulian, S.A. (1997). Infrared spectroscopy of proteins and peptides in lipid bilayers. Quarterly reviews of biophysics *30*, 365-429.

Tartaglia, G.G., and Vendruscolo, M. (2008). The Zyggregator method for predicting protein aggregation propensities. Chemical Society Reviews *37*, 1395-1401.

Tedeschi, G., Mangiagalli, M., Chmielewska, S., Lotti, M., Natalello, A., and Brocca, S. (2017). Aggregation properties of a disordered protein are tunable by pH and depend on its net charge per residue. Biochimica et Biophysica Acta (BBA)-General Subjects.

Timm, D.E., Ross, A.H., and Neet, K.E. (1994). Circular dichroism and crosslinking studies of the interaction between four neurotrophins and the

extracellular domain of the low-affinity neurotrophin receptor. Protein Science *3*, 451-458.

Tompa, P. (2002). Intrinsically unstructured proteins. Trends in biochemical sciences *27*, 527-533.

Tompa, P. (2012a). Intrinsically disordered proteins: a 10-year recap. Trends in biochemical sciences *37*, 509-516.

Tompa, P. (2012b). On the supertertiary structure of proteins. Nature chemical biology *8*, 597-600.

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

Tompa, P., Szasz, C., and Buday, L. (2005). Structural disorder throws new light on moonlighting. Trends in biochemical sciences *30*, 484-489.

Trevino, S.R., Scholtz, J.M., and Pace, C.N. (2008). Measuring and increasing protein solubility. Journal of pharmaceutical sciences *97*, 4155-4166.

Tria, G., Mertens, H.D.T., Kachala, M., and Svergun, D.I. (2015). Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. IUCrJ *2*, 207-217.

Uversky, V., and Longhi, S. (2011). Instrumental analysis of intrinsically disordered proteins: assessing structure and conformation, Vol 3 (John Wiley & Sons).

Uversky, V.N. (1993). Use of fast protein size-exclusion liquid chromatography to study the unfolding of proteins which denature through the molten globule. Biochemistry *32*, 13288-13298.

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

Uversky, V.N. (2002b). What does it mean to be natively unfolded? The FEBS Journal *269*, 2-12.

Uversky, V.N. (2009). The mysterious unfoldome: structureless, underappreciated, yet vital part of any given proteome. BioMed Research International *2010*.

Uversky, V.N. (2011). Intrinsically disordered proteins from A to Z. The international journal of biochemistry & cell biology *43*, 1090-1103.

Uversky, V.N. (2012). Size-exclusion chromatography in structural analysis of intrinsically disordered proteins. Intrinsically Disordered Protein Analysis: Volume 2, Methods and Experimental Tools, 179-194.

Uversky, V.N. (2013a). A decade and a half of protein intrinsic disorder: biology still waits for physics. Protein Science *22*, 693-724.

Uversky, V.N. (2013b). Intrinsic disorder-based protein interactions and their modulators. Current pharmaceutical design *19*, 4191-4213.

Uversky, V.N. (2013c). Unusual biophysics of intrinsically disordered proteins. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics *1834*, 932-951.

Uversky, V.N. (2014). Introduction to intrinsically disordered proteins (IDPs). (ACS Publications).

Uversky, V.N. (2017). The roles of intrinsic disorder-based liquid-liquid phase transitions in the "Dr. Jekyll–Mr. Hyde" behavior of proteins involved in amyotrophic lateral sclerosis and frontotemporal lobar degeneration. Autophagy, 1-48.

Uversky, V.N., and Dunker, A.K. (2010). Understanding protein non-folding. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics *1804*, 1231-1264.

Uversky, V.N., and Fink, A.L. (2004). Conformational constraints for amyloid fibrillation: the importance of being unfolded. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics *1698*, 131-153.

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

Uversky, V.N., Li, J., and Fink, A.L. (2001). Evidence for a partially folded intermediate in α-synuclein fibril formation. Journal of Biological Chemistry *276*, 10737-10744.

Uversky, V.N., Oldfield, C.J., and Dunker, A.K. (2005). Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. Journal of Molecular Recognition *18*, 343-384.

Uversky, V.N., Oldfield, C.J., and Dunker, A.K. (2008). Intrinsically disordered proteins in human diseases: introducing the D2 concept. Annu. Rev. Biophys. *37*, 215-246.

Uversky, V.N., Santambrogio, C., Brocca, S., and Grandori, R. (2012). Length-dependent compaction of intrinsically disordered proteins. FEBS letters *586*, 70-73.

Van Der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gsponer, J., and Jones, D.T. (2014). Classification of intrinsically disordered regions and proteins. Chemical reviews *114*, 6589-6631.

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

Weathers, E.A., Paulaitis, M.E., Woolf, T.B., and Hoh, J.H. (2004). Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. FEBS letters *576*, 348-352.

Weiss, W.F., Young, T.M., and Roberts, C.J. (2009). Principles, approaches, and challenges for predicting protein aggregation rates and shelf life. Journal of pharmaceutical sciences *98*, 1246-1277.

Wilkins, D.K., Grimshaw, S.B., Receveur, V., Dobson, C.M., Jones, J.A., and Smith, L.J. (1999). Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques. Biochemistry *38*, 16424-16431.

Williams, P.D., Pollock, D.D., and Goldstein, R.A. (2001). Evolution of functionality in lattice proteins. Journal of Molecular Graphics and Modelling *19*, 150-156.

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

Wu, H., and Fuxreiter, M. (2016). The structure and dynamics of higher-order assemblies: amyloids, signalosomes, and granules. Cell *165*, 1055-1066.

Wuttke, R., Hofmann, H., Nettels, D., Borgia, M.B., Mittal, J., Best, R.B., and Schuler, B. (2014). Temperature-dependent solvation modulates the dimensions of disordered proteins. Proceedings of the National Academy of Sciences *111*, 5213-5218.

Xie, H., Vucetic, S., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Uversky, V.N., and Obradovic, Z. (2007). Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. Journal of proteome research *6*, 1882-1898.

Xue, B., Dunbrack, R.L., Williams, R.W., Dunker, A.K., and Uversky, V.N. (2010a). PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics *1804*, 996-1010.

Xue, B., Dunker, A.K., and Uversky, V.N. (2012). Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. Journal of Biomolecular Structure and Dynamics *30*, 137-149.

Xue, B., Williams, R.W., Oldfield, C.J., Dunker, A.K., and Uversky, V.N. (2010b). Archaic chaos: intrinsically disordered proteins in Archaea. BMC Systems Biology *4*, S1.

Yakovchuk, P., Protozanova, E., and Frank-Kamenetskii, M.D. (2006). Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. Nucleic acids research *34*, 564-574.

Yang, F. (1997). The molecular structure of green fluorescent protein.

Zhang, K.Q., and Liu, X.Y. (2004). In situ observation of colloidal monolayer nucleation driven by an alternating electric field. Nature *429*, 739-743.

Zhou, H.-X. (2002). A Gaussian-chain model for treating residual charge–charge interactions in the unfolded state of proteins. Proceedings of the National Academy of Sciences *99*, 3569-3574.

BBA - General Subjects 1861 (2017) 2543–2550

Contents lists available at ScienceDirect

BBA - General Subjects

journal homepage: www.elsevier.com/locate/bbagen

Aggregation properties of a disordered protein are tunable by pH and depend on its net charge per residue

Giulia Tedeschi, Marco Mangiagalli, Sara Chmielewska, Marina Lotti, Antonino Natalello*, Stefania Brocca*

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

ARTICLE INFO

Keywords:
IDPs
Isoelectric point
NCPR
Protein solubility
Protein aggregation

ABSTRACT

Intrinsically disordered proteins (IDPs) possess a peculiar amino acid composition that makes them very soluble. Nevertheless, they can encounter aggregation in physiological and pathological contexts. In this work, we addressed the issue of how electrostatic charges can influence aggregation propensity by using the N-terminus moiety of the measles virus phosphoprotein, PNT, as a model IDP. Taking advantage of the high sequence *designability* of IDPs, we have produced an array of PNT variants sharing the same hydrophobicity, but differing in net charges per residue and isoelectric points (pI). The solubility and conformational properties of these proteins were analysed through biochemical and biophysical techniques in a wide range of pH values and compared with those of the green fluorescence protein (GFP), a globular protein with lower net charge per residue, but similar hydrophobicity. Tested proteins showed a solubility minimum close to their pI, as expected, but the pH-dependent decrease of solubility was not uniform and driven by the net charge per residue of each variant. A parallel behaviour was observed also in fusion proteins between PNT variants and GFP, which minimally contributes to the solubility of chimeras. Our data suggest that the overall solubility of a protein can be dictated by protein regions endowed with higher net charge per residue and, hence, prompter to respond to pH changes. This finding could be exploited for biotechnical purposes, such as the design of solubility/aggregation tags, and in studies aimed to clarify the pathological and physiological behaviour of IDPs.

1. Introduction

Protein aggregation is involved in a number of physiological and pathological events. Moreover, it is a major hurdle in the production and storage of recombinant proteins, included drugs. Hence, understanding the physical and chemical bases of protein aggregation could help not only to figure out how physio-pathological processes occur, but also to exploit this phenomenon for biotechnical purposes, for instance to increase *in-vitro* solubility of proteins [1], to design biomaterials with tunable aggregation properties [2], or even to design tags exploitable in the production of recombinant proteins [3].

How to recognize or predict protein solubility? Different definitions and criteria have been proposed, based on experimental observations, databases of soluble and insoluble proteins, or on the employment of machine-learning algorithms [4–6]. It emerges that besides sequence and structural features, the electrostatic properties of proteins, *i.e.* their net charge, can play a key role. It is well recognized that proteins behave as amphoteric molecules, showing significantly reduced solubility and even precipitation at their isoelectric points (pIs) [7]. On the other hand, charges can produce opposite effects. Indeed, "supercharging" of proteins, especially with negative charges, may enhance solubility [8,9], whereas positively-charged surface patches correlate with insolubility of proteins expressed in a cell-free *Escherichia coli* system [10]. Systematic studies on protein solubility find obvious limitations in the disastrous structural effects induced by extensive replacement of charged residues on globular proteins. In this context, intrinsically disordered proteins (IDPs) provide a very versatile tool to extend the "host-guest approach" [11] from peptides to larger molecules, minimizing structural effects. IDPs are usually well soluble proteins lacking strict spatial constraints and compositional complexity [12–15]. Due to their high *designability* [16], starting from a "prototypical" sequence, it is possible to generate an ideally infinite series of *ad-hoc* proteins sharing some properties (*e.g.*, hydrophobicity, length, "depth" of structural disorder, etc) and differing in others (*e.g.*, net

Abbreviations: ATR, attenuated total reflection; CD, circular dichroism (spectroscopy); FTIR, Fourier transform infrared (spectroscopy); GFP, green fluorescent protein; IMAC, immobilized-metal affinity chromatography; IDPs, intrinsically disordered proteins; FCR, fraction of charged residue; NCPR, net charge per residue; PB, phosphate buffer; PNT, N-terminus moiety of measles virus phosphoprotein; pI, isoelectric point; RC, random coil

* Corresponding authors at: Department of Biotechnology and Biosciences, State University of Milano-Bicocca, Piazza della Scienza 2, 20126 Milano, Italy.
E-mail addresses: antonino.natalello@unimib.it (A. Natalello), stefania.brocca@unimib.it (S. Brocca).

http://dx.doi.org/10.1016/j.bbagen.2017.09.002
Received 19 June 2017; Received in revised form 6 September 2017; Accepted 7 September 2017
Available online 08 September 2017
0304-4165/

charge, charge density and distribution etc.). IDPs have proved to be less prone to β-aggregation [17] and more stable to heat and pH than their folded counterparts [18,19]. These features arise from the peculiar amino acid composition of IDPs and are consistent with the abundance of highly soluble residues (proline, charged and polar residues) and the paucity of aromatic and hydrophobic residues [20–22]. These properties have suggested the use of IDPs as solubility enhancers, and the hypothesis they can act as "entropic bristles" sweeping the space around the fusion protein and preventing large molecules to participate in aggregation [23]. Nevertheless, the high solubility of IDPs does not imply a lower propensity to collective interactions, such those giving rise to aggregates or coacervates. Indeed, besides hydrophobicity, also entropic factors, hydrogen bonding and electrostatic interactions can cause aggregation [17]. Moreover, the water solubility of IDPs might depend on conformational compactness that, in its turn, is influenced by the water exposure of solubility-promoting amino acids [24]. An attempt to rationalize the relationships between electrostatic charges and conformation of IDPs is represented by charge-hydropathy plots [20] and, more recently, by diagrams of states. Through these latter empirical diagrams, conformational states of IDPs have been related to fraction of charged residue (FCR) and net charge per residue (NCPR) [25–27].

In this complex scenario, we aimed to shed light on the role of charged amino acids on IDPs solubility at pI, using as a model the N-terminus moiety of measles virus phosphoprotein (PNT) [28]. We compared at various pHs the aggregation propensity of wild-type PNT (*wt* PNT) and synthetic variants of PNT with higher net charge and markedly more acidic or basic pI. We included in our study the green fluorescent protein (GFP), a globular protein very similar to *wt* PNT in terms of net charge and pI, but differing in NCPR, which is the worthiest parameter to compare the net charge of proteins of different length. Furthermore, we explored the ability of all PNT variants and GFP to reciprocally influence their solubility in chimeric constructs.

Our study shows that overall PNTs are more pH-responsive than GFP, which has lower NCPR. Among PNT variants, the loss of solubility occurs to varying degree, depending on the protein net charge. PNT variants endowed with highest NCPR promptly undergo aggregation at or near their pI, whereas low-NCPR proteins mildly react to pH, remaining mostly soluble. We further report that PNT variants "transmit" their solubility profile to chimeric constructs with GFP. This information would greatly help in the *de-novo* design of synthetic, disordered solubility/aggregation tags and hopefully in understanding *in-vivo* processes of IDP condensation and aggregation.

2. Materials and methods

2.1. Gene design and cloning

Wild-type PNT (*wt* PNT) was cloned in pET-21a [PNT] vector [29]. Acidic and basic variants of PNT were obtained through gene synthesis (Genscript, Piscataway, NJ, USA). Two kinds of *supercharged* (*sc*) variants of PNT were designed. In the *sc-acidic* PNT, His, Lys and Arg residues of *wt* PNT were substituted with either Glu or Asp; the *basic* variant and the *sc-basic* PNT variants were obtained by substitution of Glu and Asp residues with Lys and Arg residues. Synthetic genes were cloned into pET-21a vector (EMD, Millipore, Billerica, MA, USA), between the sites *Nsi*I and *Not*I, giving rise to plasmids pET-21a [*sc-acidic* PNT], pET-21a [*basic* PNT], pET-21a [*sc-basic* PNT]. In this work, we indicate as pET-21a [PNTs] the ensemble of expression vectors carrying aforesaid PNT genes.

Constructs for the fusion of GFP at the C-terminus of PNT mutants were obtained by cloning the GFP gene into pET-21a [PNTs] digested with *Nde*I. The coding sequence was amplified by PCR from pET-19b [GFP] [29] with primers inserting *Nde*I restriction sites at both 5′ and 3′ extremities. The forward and reverse primers for amplification were: FW 5′-GGATCCCATATGAAAGTGAGCAAG-3′, RV 5′-CATATGCCCAAGCTTCTTGTACAG -3′ (*Nde*I restriction site is underlined). Amplification reactions were carried out using Q5® High-Fidelity DNA Polymerase (New England Biolabs, Ipswich, MA). The reaction conditions used were: 1 cycle (98 °C for 2 min), 25 cycles (98 °C 10 s, 56 °C 30 s, and 72 °C 1 min), and a final cycle of 72 °C 3 min. The PCR product was preliminarily cloned into pUC18 blunt-end digested with *Sma*I obtaining pUC18 [GFP]. The GFP gene was then excised from pUC18 [GFP] digested with *Nde*I and gel-purified before ligation into the pET-21a [PNTs] cleaved with the same restriction enzyme.

The correct orientation of the GFP insert in the pET-21a [PNTs-GFP] vectors was verified by enzyme restriction and by bidirectional DNA sequencing. The amino acid sequences of PNTs are reported in Fig. S1. GFP was produced from pET-19b [GFP] [29].

2.2. Protein production and purification

Escherichia coli strain BL21[DE3] (EMD, Millipore, Billerica, MA, USA) was used as the host for heterologous production of PNTs variants. Transformed cells were grown overnight at 37 °C in Lennox medium (10 g/L tryptone, 5 g/L yeast extract, 5 g/L NaCl), diluted 1:20 in 200 mL of Zym-5052 medium [30] and incubated at 25 °C. Media were added of 100 mg/L ampicillin.

Proteins were extracted as described in [31] and recombinant, his-tagged proteins were purified by immobilized-metal affinity chromatography (IMAC) on Ni/NTA agarose gel (Jena Bioscience, Jena, Germany) at 4 °C. To improve the purification yield, clarified lysates were incubated at 4 °C for 1 h with Ni/NTA agarose gel before purification.

Protein concentration was determined by the Bradford protein assay (Bio-Rad, California, USA), using bovine serum albumin as a standard.

Samples containing highest protein concentrations were buffer exchanged twice by gel filtration on PD10 column (GE Healthcare, Little Chalfont, UK) against 10 mM ammonium acetate buffer pH 7.0.

2.3. Biochemical and biophysical analyses

Since pH strongly impacts on protein solubility and affects determination of protein concentration by Bradford assay, samples were prepared by a procedure allowing to minimize differences in protein yield and sample concentration. After buffer exchange in 10 mM ammonium acetate, elution fractions containing protein at the highest concentrations were pooled, newly quantified and divided in samples containing the same protein amount. Samples were lyophilized in a freeze-dryer (Heto FD1.0 Gemini BV, Apeldoorn, Netherlands) and then suspended in equal volumes of 10 mM potassium phosphate buffer (PB) at different pH values (3.0, 5.0, 6.0, 7.0, 9.0). Only GFP, PNT *basic* and PNT *basic*-GFP were assayed also at pH 8.0, 8.5, 9.5, 10.0 and 11.0; while *sc-acidic* PNT were further assayed at pH 4.0. The pH measurements were carried out at room temperature with a HI 9321 Microprocessor pH meter (Hanna Instruments, Italy). The instrument was calibrated against the standard pH 4.00 and 7.00 solutions (Sigma Aldrich, St. Louis, MO, USA).

2.3.1. Far-UV circular dichroism (CD) spectroscopy

Lyophilized samples were suspended in PB (0.09 mg/ml for GFP and PNTs variants and 0.18 mg/ml for fusion proteins) at different pH values, and incubated for 1 h at room temperature. CD spectra were recorded at room temperature by a spectropolarimeter J-815 (JASCO Corporation, Easton, USA) in a 1-mm path-length cuvette. Measurements were performed at variable wavelength (190–260 nm) with scanning velocity 20 nm/min, bandwidth 1 nm, digital integration time per data 2 s and data pitch 0.2 nm. All spectra were averaged from two independent acquisitions, corrected for buffer contribution, and smoothed by Means-Movement algorithm. Experiments were performed in triplicate.

2.3.2. Fourier transform infrared (FTIR) spectroscopy

Lyophilized samples were suspended in PB (1.5 mg/ml) at different pH values and incubated for 1 h at room temperature. Two microliters of the above protein solutions were deposed on the single reflection diamond element of the attenuated total reflection (ATR) device (Quest, Specac, USA) and dried at room temperature to obtain a protein film [32,33]. The protein film on the ATR element was hydrated by adding 6 μL of D_2O close to the sample [33,34] and incubated for 1 h at room temperature. The ATR/FTIR spectra were collected at room temperature using a Varian 670-IR spectrometer (Varian Australia Pty Ltd., Mulgrave, Victoria, Australia), equipped with a nitrogen-cooled Mercury Cadmium Telluride detector, under the following conditions: resolution 2 cm^{-1}, scan speed 25 kHz, 1000 scan co-additions, triangular apodization, and dry-air purging.

ATR/FTIR absorption spectra were corrected for buffer contribution, normalized at the Amide *I′* (1700–1600 cm^{-1}) band area and were smoothed using the Savitsky-Golay method before second derivative calculation. Spectral analyses were performed with the Resolutions-Pro software (Varian Australia Pty Ltd., Mulgrave, Victoria, Australia). At least three independent measurements were performed for each condition.

2.3.3. Solubility assay

SDS-PAGE was used to assess solubility of GFP and PNT variants after incubation at different pH. Lyophilized samples were suspended in PB at different pH values, at concentration of 0.5 mg/mL. After 1-h incubation, an aliquot was collected (total protein), and the remaining was centrifuged for 10 min at 15,000 × *g* to separate soluble and insoluble protein fractions. An equal volume (20 μl) of total and soluble proteins were separated in 14% SDS-PAGE [35] and stained with GelCode Blue (Pierce, Rockford, USA). Broad-range, pre-stained molecular-weight markers (GeneSpin, Milan, Italy) were used as standards. Densitometric volume of each protein band was calculated by the software Image Lab (Bio-Rad, California, USA). For each pH value, the relative amount of soluble protein (solubility) was calculated with reference to total protein in the aliquot. Percentages are referred to the highest value of solubility considered as 100%. Data represent an average of three independent biological replicates. Similar results were obtained from solubility tests carried out after 30 min-, 1 h- and 2 h-incubation for PNT variants and for GFP, at pH near the pI of each protein (data not shown).

2.4. Bioinformatic analysis

The theoretical pI was calculated with different algorithms: Expasy ProtParam (http://web.expasy.org/protparam) and Isoelectric point calculator [36]. Disorder prediction with Pondr-fit [37] and plots of mean net charge *versus* mean hydropathy [20] were used to assess conformation profile.

NCPR values were calculated as:

$$NCPR = \frac{(aa\ positive - aa\ negative)}{aa\ total}$$

FCR values were calculated as:

$$FCR = \frac{(aa\ positive) + (aa\ negative)}{aa\ total}$$

where *aa positive* is the number of positively charged amino acids, *aa negative* is the number of negative charged amino acids and *aa total* is the total number of amino acids [25].

NCPR, FCR and the Kyte-Doolittle hydropathy score (scaled from 0, least hydrophobic, to 9, most hydrophobic) were calculated through the webserver CIDER [38].

3. Results and discussion

Studies on aggregation/solubility of proteins are very challenging if we consider the faceted role different amino acid residues can have, depending on their physicochemical classes, solvent exposition, and on their position in a protein structure [39]. Although IDPs have to be considered as conformational ensembles, their use as a model allows to greatly simplify the issue, as it allows to reduce the relevance of conformational effects and to focus on the “chemical behaviour” of “biological objects”. Moreover, the relaxed conformational constrains on IDPs primary structure made it possible to design a “family” of related proteins that can be assimilated to ionisable amphoteric polyelectrolytes, whose response to chemical and physical laws can be gathered more easily than from a single protein.

3.1. Design of PNT variants and of their fusions with the green fluorescent protein (GFP)

To study systematically the solubility of a disordered protein, PNT variants were designed exploring a wide range of pI values and net charges. More in detail, our experimental approach was aimed at sampling two mild-charged and two *supercharged* (*sc*) basic and acidic variants of PNT. Since *wt* PNT already exhibits mild-acidic features, we designed three synthetic variants, thereof one is mild basic (simply referred as *basic*), one *sc-acidic* and one *sc-basic*. In the following, the ensemble of PNT variants used in this work is referred to as “PNTs”. Overall, the design of synthetic PNT variants was carried out by reversing the sign of charged residues already present in the wild-type sequence while keeping unchanged all other residues (Fig. 1A). For this reason, all PNTs have a very similar fraction of charged residues (FCR, 0.257 ± 0.004) and hydropathy score (3.826 ± 0,067), as calculated by CIDER [38], and as shown in the Uversky plot [20] (Fig. 1B).

Wild-type PNT was described previously [29]. Briefly, it spans the first 230 amino acid residues of the whole P protein and carries an N-terminal 6xHis tag and a C-terminal tail containing the TEV protease target sequence. This sequence contains 21 positively charged amino acid residues and 38 negatively charged residues resulting in an acidic pI of 4.88 and featuring an NCPR of - 0.071. The mild-charged *basic* variant was obtained by almost inverting the ratio of positively (37) and negatively (23) charged amino acid residues of *wt* PNT, and reaching a pI of 9.61 and NCPR of + 0.055. The *sc-acidic* PNT has a pI of 3.37 and includes 62 negatively charged residues (0 positives ones), with an NCPR of - 0.248, while *sc-basic* PNT has a pI of 11.44 and includes 57 positively charged residues (0 negatives ones), with an NCPR of + 0.216. We assumed that the 6xHis tag and TEV site affect all variants in the same way, producing effects negligible in the comparative analyses. The features of PNTs are summarized in Table 1, amino acid sequences are reported in Fig. S1 and plots of linear net charge per residue in Fig. S2. Despite the profound sequence changes so far described, the overall disorder profile of synthetic PNTs calculated by Pondr-fit [37] remains similar to that of *wt* PNT, and slightly more disordered for the two *sc*-PNTs (Fig. 1C).

Each PNT variant was C-terminally fused to GFP to assay the ability of each moiety to affect the solubility of the fusion partner. The GFP shares with PNTs a very similar hydropathy score (3.94) and FCR (0.246), but has lower NCPR (− 0.023). Features of PNTs fused with GFP are included in Table 1.

All proteins but *sc-basic* PNT and *sc-basic* PNT-GFP were produced in Zym 5052 medium and purified at comparable yield (~4 mg per liter of colture) from the soluble fraction of cell extracts. In the case of *sc-basic* PNT and *sc-basic* PNT-GFP, we did not observe any production of the recombinant proteins, even in the insoluble protein fraction. This problem has been already referred by other Authors for a supercharged globular protein [40]. We can hypothesize that the high frequency of Lys and Arg residues in *sc-basic* PNT sequence may unfavourably impact on its translation rate, hence producing ribosome stalling and transcript

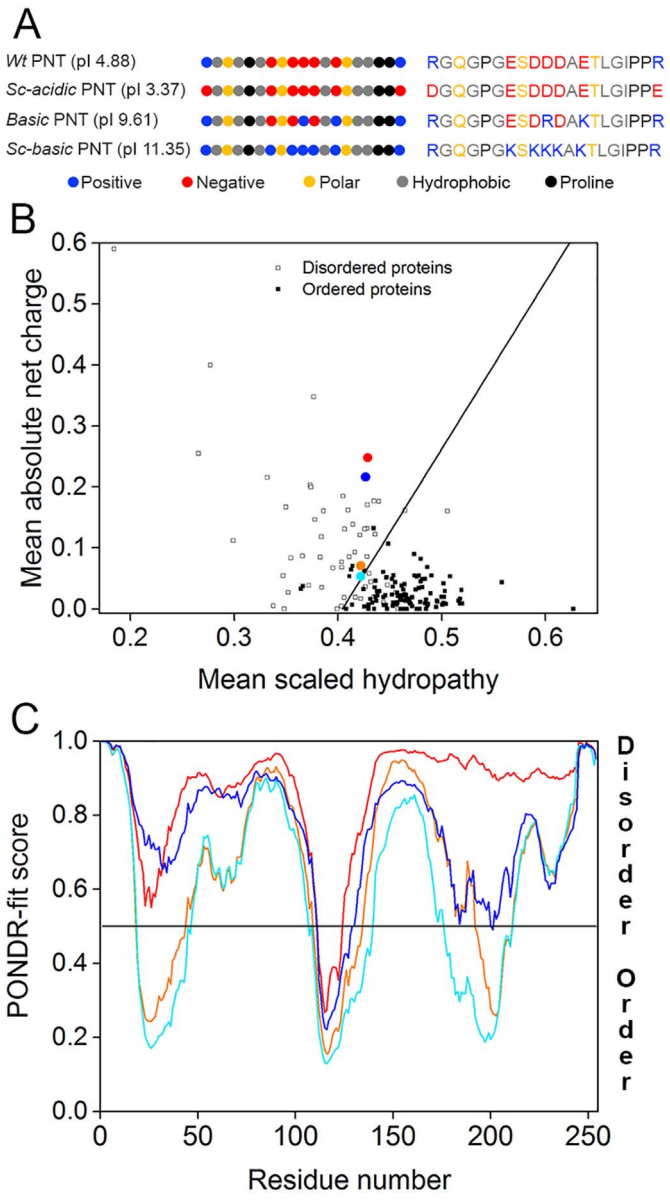


Fig. 1. Design and disorder prediction of PNT variants. (A) Scheme of amino acid composition of PNT variants. Residues 85–105 are shown to exemplify the results obtained substituting charged residues in the same relative positions and keeping unchanged polar, hydrophobic and proline residues. (B–C) Predictions were carried out using charge-hydropathy [20] (B) and Pondr fit [37] (C) predictors. *Wt* PNT, *sc-acidic* PNT, *basic* PNT and *sc-basic* PNT are indicated in orange, red, light blue and blue, respectively.

Table 1
Features of proteins assayed in this work. The amino acid sequences are reported in S1 and include 6xHis tag and TEV site. Along this paper, we will simply refer to the mean value of pI.

Protein ID	Amino acid content					pI	NCPR
	Lys	Arg	His	Glu	Asp		
Wt PNT	9	12	11	23	16	4.88 ± 0.03	− 0.071
Sc-acidic PNT	–	–	8	34	29	3.37 ± 0.06	− 0.248
Basic PNT	16	21	11	15	8	9.61 ± 0.37	+ 0.055
Sc-Basic PNT	45	12	11	–	–	11.35 ± 0.15	+ 0.216
Wt PNT-GFP	31	18	21	39	35	5.24 ± 0.06	− 0.048
Sc-Acidic PNT-GFP	22	6	18	50	48	4.16 ± 0.05	− 0.139
Basic PNT-GFP	38	27	21	31	27	8.45 ± 0.40	+ 0.014
Sc-Basic PNT-GFP	67	18	21	16	19	10.20 ± 0.25	+ 0.099
GFP	22	6	10	16	19	6.15 ± 0.09	− 0.023

degradation [41]. Further attempts to produce other *sc-basic* PNTs with slightly modified sequences were unsuccessful. Since *sc-acidic* PNT, *wt* PNT and *basic* PNT allow us to sample a wide range of NCPR and pI values, we considered this ensemble of proteins, along with GFP, suitable to test our hypothesis. All proteins were purified by IMAC, liophilyzed, and resuspendend in phosphate buffer (PB) adjusted at different pH and finally solubility was assessed. All samples analysed before and after lyophilisation gave superimposable spectra of far-UV CD and FTIR spectroscopies (data not shown).

3.2. Solubility and propensity to aggregation of PNT variants and GFP

The solubility at different pHs of PNT variants and GFP was studied *in vitro* using three complementary techniques: solubility assays, far-UV CD and FTIR spectroscopies. We have considered a “standard range” of pH values (3.0, 5.0, 6.0, 7.0, 9.0) to analyze all the proteins and compare at a glance their solubility profiles. Other pH values were chosen *ad hoc* to study more extensively some of the proteins (see later).

The CD spectrum of *wt* PNT at pH 7.0 is that typical of a disordered protein, with a deep downward peak in the range 190–200 nm (Fig. 2.A). The shape of this spectrum is consistent with that already published for the same protein and measured in sodium phosphate buffer at pH 7.5 [29]. The ellipticity value observed at 222 nm is consistent with the existence of some residual helical structure. Overall, at pH 7.0, PNTs spectra are similar as for profile and ellipticity. As the pH reaches the pI value of each PNT variant, we observed a dramatic loss of the ellipticity signals (Fig. 2.A–C). Moreover, in the *sc-acidic* PNT sample we detected an increase of ellipticity at 190 nm and a shift of the minimum toward 218 nm, suggesting the simultaneous formation of β-structure. Overall, far-UV CD spectroscopy analyses hint that PNTs undergo aggregation as pH approaches to their pI. Solubility assay and FTIR analyses were performed to assess this hypothesis.

Solubility was quantified by densitometric analysis of samples after SDS-PAGE separation. We detected the lowest solubility of *wt*, *basic* and *sc-acidic* PNT at pH 5.0, 3.0 and 9.0, respectively (Fig. 2.D–F). This observation is in good agreement with the flattening of CD signal observed under the same conditions. It is worth to notice that the decrease in measured solubility is higher for *sc-acidic* PNT (∼ − 95%) than for *wt* (∼ − 60%) and *basic* PNT (∼ − 50%).

The FTIR second derivative spectra (whose minima correspond to absorption maxima) of PNTs were reported in the Amide *I′* band in Fig. 2.G–I. We show here spectra obtained after H/D exchange, since they allow to better resolve the spectral signature of different structural secondary elements and to distinguish between α-helical and disordered structures.

A scheme of the typical absorption regions of the different protein secondary structures for samples in D_2O is explicitly reported in the spectrum of *wt* PNT in Fig. 2.G [42,43]. The FTIR second derivative spectra of PNTs at pH 7.0 show a main component around 1641 cm^{-1} (Fig. 2.G–I) that can be assigned to disordered structures.

According to solubility assays, spectra collected at different pHs show an additional component around 1619–1613 cm^{-1} (arrow in Fig. 2.G–I), whose intensity increases as pH reaches the pIs of the different PNTs and indicates the formation of intermolecular β-sheets [42,43].

For the sake of completeness, we also measured the solubility of GFP at different pHs. GFP is a globular protein endowed with a well-defined β-barrel structure composed of 11 β-strands [44], and a theoretical pI of 6.15. CD spectra between pH 6.0 and 9.0 show a positive peak at 195 nm and a broad negative peak at 218 nm, as expected for a natively structured protein with a predominant content of β-strands (Fig. 3.A). At pH 5.0, GFP reaches its lowest solubility, with a moderate loss of the CD signal and comparable loss of soluble protein (∼ − 20%) (Fig. 3.A, B). The difference between the observed pH dependence and the theoretical pI of GFP may be due to pKa shifts of titratable residues, which, in turn, may depend on their positions in the *core* of a folded protein

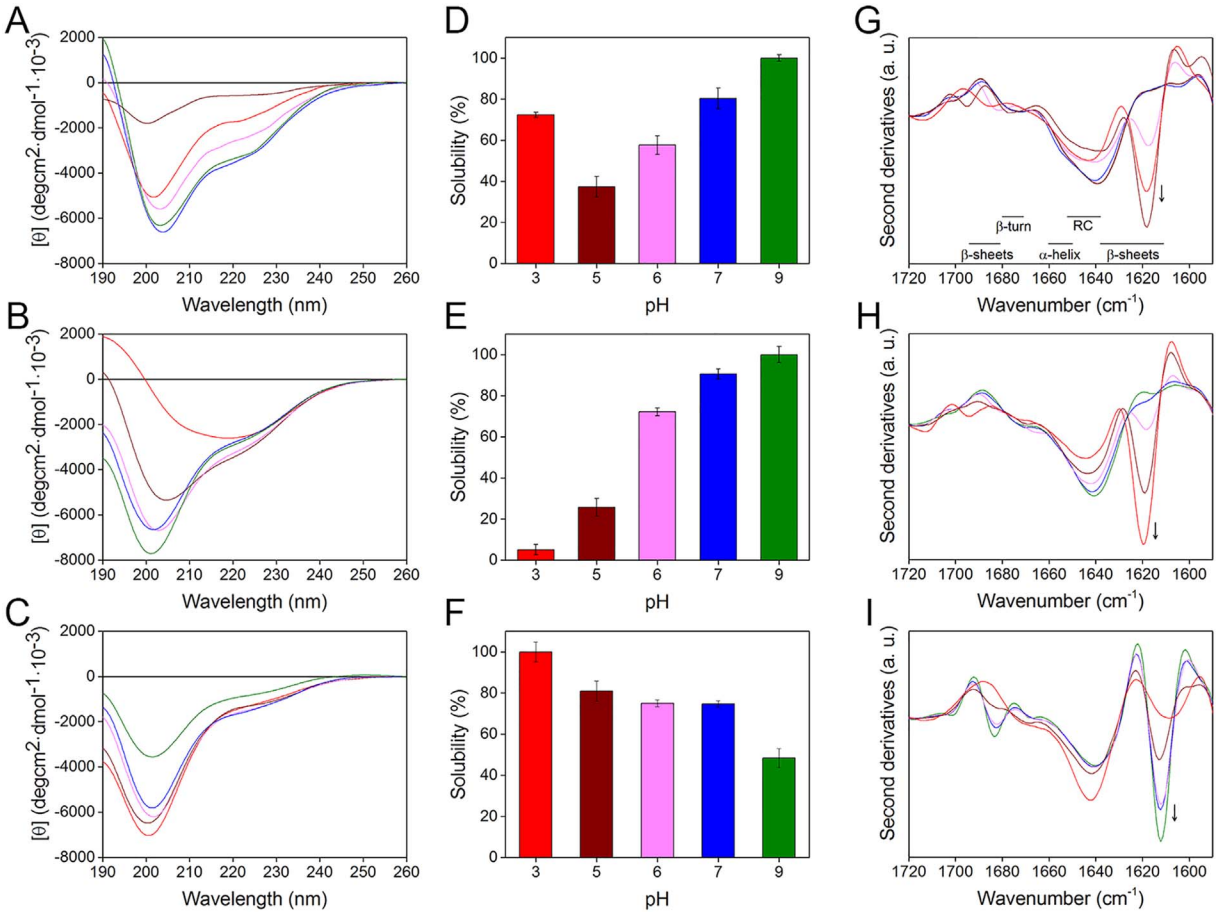


Fig. 2. Solubility and propensity to aggregation of single PNTs. For each analysis, proteins were prepared in PB at pH 3.0 (red), 5.0 (brown), 6.0 (pink), 7.0 (blue) and 9.0 (green). Upper, middle and lower row refers to *wt*, *sc-acidic* and *basic* PNT, respectively. A–C) Far-UV CD spectra. It is shown one of three independent experiments. D–F) Solubility assay. Error bars indicate standard deviations on three independent experiments. G–I) Second derivatives of the FTIR absorption spectra. Arrows point to increasing intensity of the intermolecular β-sheet peak. The Amide *I′* band assignment to the protein secondary structures is also given in G. RC: random coil. It is shown one of three independent experiments.

[39]. At pH 3.0, GFP is partially unstructured (Fig. 3.A) and yet soluble (Fig. 3.B). The structural transitions of GFP were confirmed and completed by FTIR analyses (Fig. 3.C). The second derivatives of the IR absorption spectra at pHs 6.0–9.0 show a main component at $\sim 1623\ cm^{-1}$ that, along with the peak around $1689\ cm^{-1}$, is due to native intramolecular β-sheets. At lower pHs a partial loss of the native components indicates protein unfolding, which is more evident at pH 3.0 (Fig. 3.C).

Taken together, these results highlight that changes of pHs produce a stronger impact on the solubility of PNT variants than on the solubility of the globular GFP. What makes the difference in the behaviour of GFP and PNTs? It is well reasonable to assume that compaction and folding may influence protein solubility through the exposure at different extent of solubility-promoting residues [45]. However, we considered that other protein features might be of relevance, in particular the difference in the net charge per residue that is described by NCPR [25]. When challenged at different pHs, our set of proteins aggregate at or near their pI, with different intensities which reflect the absolute value of NCPR ($NCPR^{sc\ acidicPNT} = |-0.248| > NCPR^{wtPNT} = |-0.071| > NCPR^{basicPNT} = |+0.055| > NCPR^{GFP} = |-0.023|$). Among PNTs, we observed the strongest pH-dependent aggregation with *sc-acidic* PNT, whereas the loss of solubility of *basic* PNT was the mildest. According with the reported results, we concluded that NCPR should be taken into careful consideration to

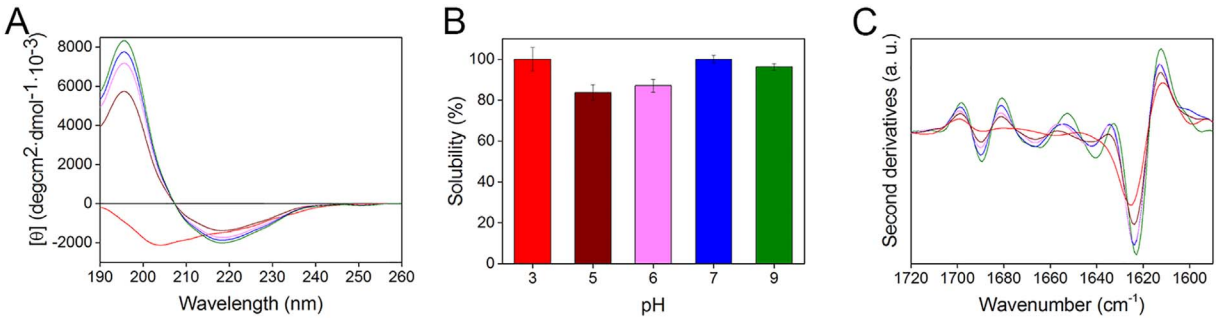


Fig. 3. Solubility and propensity to aggregation of GFP. For each analysis, GFP was in PB at pH 3.0 (red), 5.0 (brown), 6.0 (pink), 7.0 (blue) and 9.0 (green). A) Far-UV CD spectra. It is shown one of three independent experiments. B) Solubility assay. Error bars indicate standard deviations on three independent experiments. C) Second derivatives of the FTIR absorption spectra. It is shown one of three independent experiments.

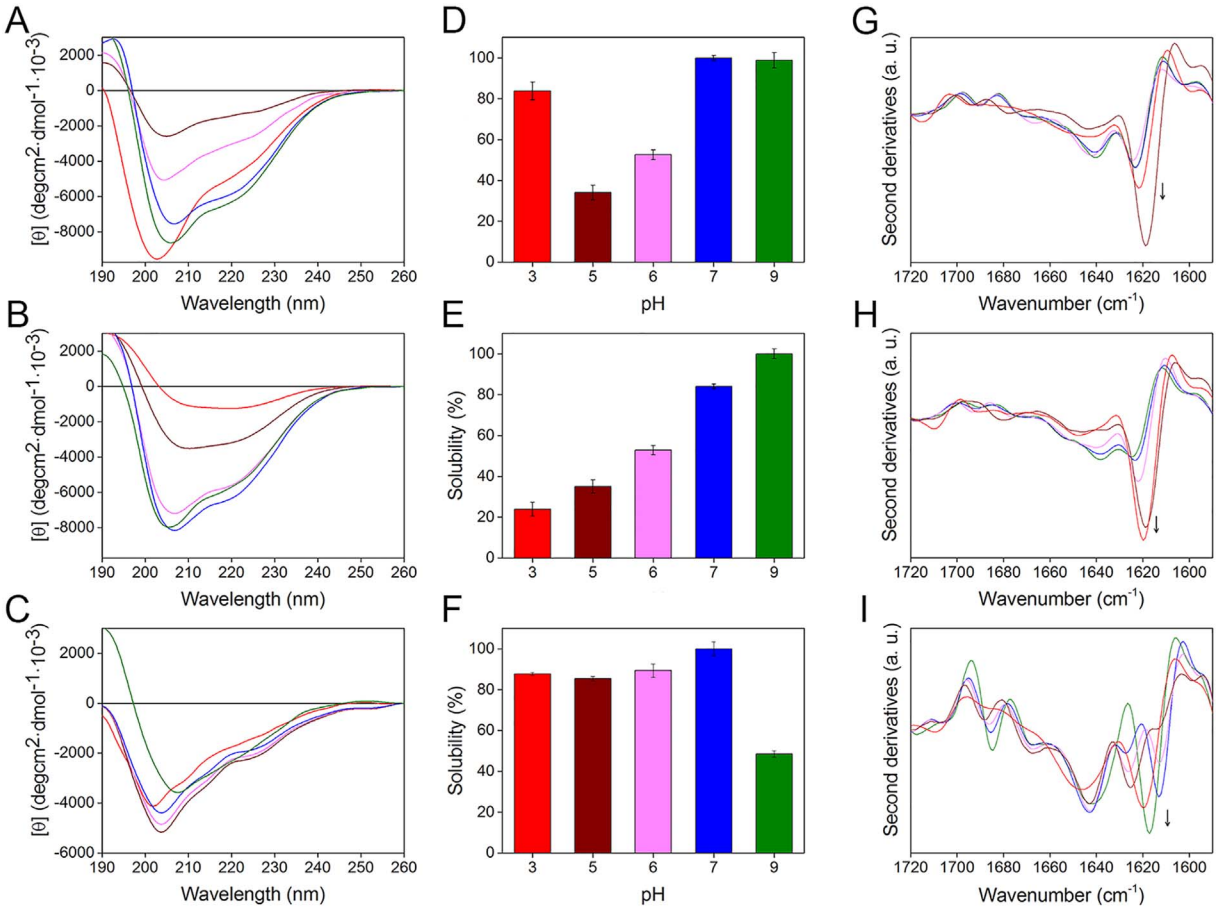


Fig. 4. Solubility and aggregation of GFP fused to PNT variants. For each analysis, proteins were in PB at pH 3.0 (red), 5.0 (brown), 6.0 (pink), 7.0 (blue) and 9.0 (green). Upper, middle and lower row refers to *wt*, *sc-acidic and basic* PNT, respectively. A–C) Far-UV CD spectra. It is shown one of three independent experiments. D–F) Solubility assay. Error bars indicate standard deviations on three independent experiments. G–I) Second derivatives of the FTIR absorption spectra. Arrows point to increasing intensity of the intermolecular β-sheet peak. It is shown one of three independent experiments.

predict pH-dependent aggregation. This interpretation is indirectly corroborated by experimental data on the high and pI-independent solubility in the pH range 2–12 of charge-free proteins, which obviously exhibit null NCPR [46].

3.3. Solubility and aggregation of GFP fused to PNT variants

To investigate the behaviour of PNTs as solubility tags, we performed the same experiments described above with chimeric proteins composed by PNT variants and GFP (Table 1). The CD spectra at pH 7.0 of all chimeras (Fig. 4.A–C) are similar to those already observed for *wt* PNT-GFP in similar conditions (PB at pH 7.5) [29], with a negative peak at 205 nm, instead of the deep downward peak typically observed around 190–200 nm in disordered proteins. When pH approaches the pI of the respective IDP moieties, a marked spectral flattening occurs. This observation is consistent with solubility profiles, which roughly parallel those observed for respective individual PNTs in the same pH range (Fig. 4.D–F). It is worth to remark that *sc-acidic* PNT-GFP undergoes the most intense loss of solubility (~ − 95%) at pH 4.0 (data not shown), likely reflecting the pI of the chimeric protein (4.16), rather than the pI of the lone PNT moiety (3.37). Such a pI shift is hard to be experimentally detected in *wt* PNT and its GFP fusion because of the proximity of their pIs (4.88 and 5.24, respectively). The behaviours of *basic* PNT and its GFP-fusion were similar even in the range of pH 8.0–11.0 (Fig. 5). Although it is difficult to generalize, we can consider that small pI differences are hardly detectable in a shallow solubility profile, as that of *basic* PNT, *vice versa* they are strikingly evident in systems that are more pH-sensitive, as that of *sc-acidic* PNT.

The FTIR second derivative spectra of the GFP fusions at pH 7.0 (Fig. 4.G–I) mainly show the sum of spectral components observed for the isolated GFP and PNT variants at the same pH. FTIR spectra are consistent with the data on solubility, since the intensity of the intermolecular β-sheet component (~1619–1613 cm^{-1}) increases as the pH approaches the pI of the disordered moiety (Fig. 4.G–I).

Since solubility/aggregation profiles of single PNTs and their GFP-fused counterparts are very similar, one can reason that GFP exerts a marginal effect on the overall solubility of chimeric proteins.

From our results, we can infer that PNT variants with the highest NCPR, *i.e. wt* and *sc-acidic* variants, are able to prime the aggregation burst of whole chimeric constructs, suggesting the use of similar IDPs as aggregation tags rather than as solubility tags. *Vice versa*, milder charged polyampholytes, *i.e. basic* PNT, are less sensitive to pH changes and can cope with a broad range of pHs without undergoing aggregation.

This information may help to better define and to rationalize the properties of an effective solubility enhancer, already described in the pivotal work of Santner et al. 2012 as an entropic bristle of similar size and different pI than the target protein [23]. Overall, our results indicate that the use of supercharged proteins as solubility enhancers is inherently risky, since high net charge, besides driving extremely high solubility, can also lead to extensive aggregation. Moreover, data reported suggest that each moiety of a fusion protein may "sense" environmental pH according to its own features. When NCPR is uneven along a sequence, "local" values of NCPR should be considered instead of a whole NCPR score, averaged on the entire sequence. For instance, the low NCPR value of *basic* PNT-GFP (+ 0.014) would induce to

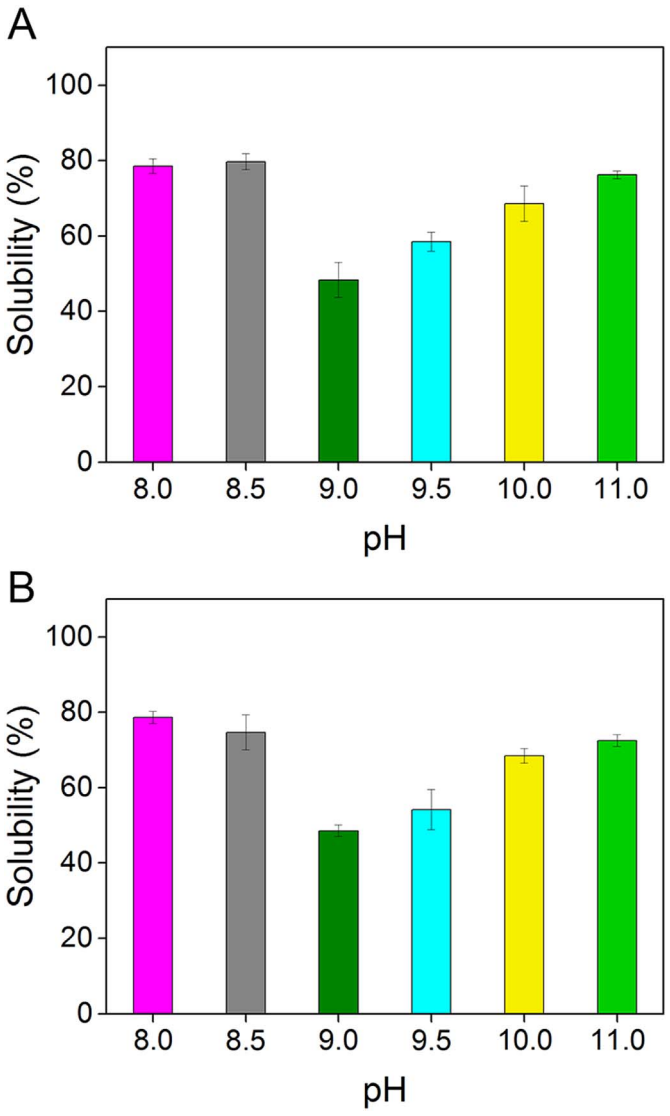


Fig. 5. Solubility assay of *basic* PNT *per se* (A) and fused to GFP (B). For each analysis, proteins were dissolved in PB at different pHs. Error bars indicate standard deviations on three independent experiments.

underestimate its propensity to aggregate driven by the disordered partner (NCPR = + 0.055). Plots of linear net charge per residue can be useful to see at a glance NCPR distribution along a sequence (see Fig. S2).

Can we generalize the behaviour observed for GFP and PNT to other globular and disordered proteins? We can reason that pH-sensitivity can be exacerbated in protein regions where one type of charged residue is recurrent (*e.g.* arginine-rich protamines; Glu/Asp-rich prothymosin α). Such low-complexity sequences are thought to enable IDPs to undergo fast, collective interactions [47,48]. Condensation of protein-rich assemblies has been recognized to foster liquid-liquid phase transitions giving rise to functionally important, membrane-less subcellular compartments, such as nucleoli, RNA granules, Cajal bodies. It is conceivable that different pH-sensitivity may impart different aggregation propensity and "phase behaviour", in response to even subtle changes of intracellular pH or NCPR. Indeed, transitions from expanded coil to collapsed globules often occur suddenly and can be reversed by even small changes in the net charge per residues [26,27], suggesting the existence of a threshold value of NCPR which delimits the two conformational ensembles. To conclude, it seems that we can still learn a lot by reconsidering and applying long-time known chemical-physical principles to new questions, such as the aggregation/coacervation of disordered proteins. The high designability of IDPs will help to experimentally prove and further understand mechanisms that may in general influence the aggregation of proteins.

4. Conclusions

We found that aggregation propensity in a set of model proteins mainly responds to pH changes according to NCPR absolute value. Besides the expected loss of solubility at pI, we found that "aggregation intensity" is directly proportional to NCPR, which correlates net charge to protein size. This implies that proteins endowed with similar net charge and pI can behave differently in terms of "aggregation intensity", according to their NCPR. Moreover, protein regions with highest NCPR leads the overall behaviour in chimeric proteins.

The overall rules dictating the aggregation appear captivating in their simplicity, in spite of the complexity of physiological and pathological phenomena in which might be involved. These observations may contribute to understand the behaviour of IDPs in response to events (*e.g.*, post-translational modifications, environment pH changes, mutations) that can affect protein NCPR. Moreover, this knowledge can have applicative potential in the design of solubility/aggregation tags for recombinant proteins.

Transparency document

The http://dx.doi.org/10.1016/j.bbagen.2017.09.002 associated with this article can be found, in online version.

Acknowledgments

This work was partly supported by a grant Fondo di Ateneo (FA) of the University of Milano-Bicocca (2016-ATE-0504) to SB, AN and ML.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.bbagen.2017.09.002.

References

[1] S.R. Trevino, J.M. Scholtz, C.N. Pace, Measuring and increasing protein solubility, J. Pharm. Sci. 97 (2008) 4155–4166.
[2] J.R. Simon, N.J. Carroll, M. Rubinstein, A. Chilkoti, G.P. López, Programming molecular self-assembly of intrinsically disordered proteins containing sequences of low complexity, Nat. Chem. (2017) 509–515.
[3] S. Costa, A. Almeida, A. Castro, L. Domingues, Fusion Tags for Protein Solubility, Purification, and Immunogenicity in *Escherichia coli*: the novel Fh8 system, Recombinant Protein Expression in Microbial Systems, 63 (2014).
[4] F. Agostini, M. Vendruscolo, G.G. Tartaglia, Sequence-based prediction of protein solubility, J. Mol. Biol. 421 (2012) 237–241.
[5] W.F. Weiss, T.M. Young, C.J. Roberts, Principles, approaches, and challenges for predicting protein aggregation rates and shelf life, J. Pharm. Sci. 98 (2009) 1246–1277.
[6] T. Niwa, B.-W. Ying, K. Saito, W. Jin, S. Takada, T. Ueda, H. Taguchi, Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of *Escherichia coli* proteins, Proc. Natl. Acad. Sci. 106 (2009) 4201–4206.
[7] J. Loeb, Amphoteric colloids ii. Volumetric analysis of ion-protein compounds; the significance of the isoelectric point for the purification of amphoteric colloids, J. Gen. Physiol. 1 (1918) 237–254.
[8] K.Q. Zhang, X.Y. Liu, In situ observation of colloidal monolayer nucleation driven by an alternating electric field, Nature 429 (2004) 739–743.
[9] Y. Su, Z. Zou, S. Feng, P. Zhou, L. Cao, The acidity of protein fusion partners predominantly determines the efficacy to improve the solubility of the target proteins expressed in *Escherichia coli*, J. Biotechnol. 129 (2007) 373–382.
[10] P. Chan, R.A. Curtis, J. Warwicker, Soluble expression of proteins correlates with a lack of positively-charged surface, Sci Rep 3 (2013) 3333.
[11] K.R. Shoemaker, P.S. Kim, E.J. York, J.M. Stewart, R.L. Baldwin, Tests of the Helix Dipole Model for Stabilization of α-Helices, (1987), pp. 563–567.
[12] V.N. Uversky, Unusual biophysics of intrinsically disordered proteins, Biochim. Biophys. Acta Protein Proteomics 1834 (2013) 932–951.
[13] A.K. Dunker, J.D. Lawson, C.J. Brown, R.M. Williams, P. Romero, J.S. Oh, C.J. Oldfield, A.M. Campen, C.R. Ratliff, K.W. Hipps, J. Ausio, M.S. Nissen,

R. Reeves, C.H. Kang, C.R. Kissinger, R.W. Bailey, M.D. Griswold, M. Chiu, E.C. Garner, Z. Obradovic, Intrinsically disordered protein, J. Mol. Graph. Model. 19 (2001) 26–59.
[14] H. Li, R. Helling, C. Tang, N. Wingreen, Emergence of preferred structures in a simple model of protein folding, Science 273 (1996) 666–669.
[15] P.D. Williams, D.D. Pollock, R.A. Goldstein, Evolution of functionality in lattice proteins, J. Mol. Graph. Model. 19 (2001) 150–156.
[16] A.K. Dunker, M.S. Cortese, P. Romero, L.M. Iakoucheva, V.N. Uversky, Flexible nets, FEBS J. 272 (2005) 5129–5148.
[17] R. Linding, J. Schymkowitz, F. Rousseau, F. Diella, L. Serrano, A comparative study of the relationship between protein structure and β-aggregation in globular and intrinsically disordered proteins, J. Mol. Biol. 342 (2004) 345–353.
[18] V. Csizmók, E. Szőllősi, P. Friedrich, P. Tompa, A novel two-dimensional electrophoresis technique for the identification of intrinsically unstructured proteins, Mol. Cell. Proteomics 5 (2006) 265–273.
[19] F. Campos, G. Guillén, J.L. Reyes, A.A. Covarrubias, A general method of protein purification for recombinant unstructured non-acidic proteins, Protein Expr. Purif. 80 (2011) 47–51.
[20] V.N. Uversky, J.R. Gillespie, A.L. Fink, Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins: Struct., Funct., Bioinf. 41 (2000) 415–427.
[21] E.A. Weathers, M.E. Paulaitis, T.B. Woolf, J.H. Hoh, Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein, FEBS Lett. 576 (2004) 348–352.
[22] S. Lise, D.T. Jones, Sequence patterns associated with disordered regions in proteins, Proteins: Struct., Funct., Bioinf. 58 (2005) 144–150.
[23] A.A. Santner, C.H. Croy, F.H. Vasanwala, V.N. Uversky, Y.-Y.J. Van, A.K. Dunker, Sweeping away protein aggregation with entropic bristles: intrinsically disordered protein fusions enhance soluble expression, Biochemistry 51 (2012) 7250–7262.
[24] R. Van Der Lee, M. Buljan, B. Lang, R.J. Weatheritt, G.W. Daughdrill, A.K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D.T. Jones, Classification of intrinsically disordered regions and proteins, Chem. Rev. 114 (2014) 6589–6631.
[25] A.H. Mao, S.L. Crick, A. Vitalis, C.L. Chicoine, R.V. Pappu, Net charge per residue modulates conformational ensembles of intrinsically disordered proteins, Proc. Natl. Acad. Sci. 107 (2010) 8183–8188.
[26] R.K. Das, R.V. Pappu, Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues, Proc. Natl. Acad. Sci. 110 (2013) 13392–13397.
[27] R.K. Das, K.M. Ruff, R.V. Pappu, Relating sequence encoded information to form and function of intrinsically disordered proteins, Curr. Opin. Struct. Biol. 32 (2015) 102–112.
[28] D. Karlin, S. Longhi, V. Receveur, B. Canard, The *N*-terminal domain of the phosphoprotein of morbilliviruses belongs to the natively unfolded class of proteins, Virology 296 (2002) 251–262.
[29] I. Sambi, P. Gatti-Lafranconi, S. Longhi, M. Lotti, How disorder influences order and vice versa - mutual effects in fusion proteins containing an intrinsically disordered and a globular protein, FEBS J. 277 (2010) 4438–4451.
[30] F.W. Studier, Protein production by auto-induction in high-density shaking cultures, Protein Expr. Purif. 41 (2005) 207–234.
[31] F. Parravicini, S. Brocca, M. Lotti, Evaluation of the conformational stability of recombinant desulfurizing enzymes from a newly isolated *Rhodococcus* sp, Mol. Biotechnol. (2015) 1–11.
[32] F. Parravicini, A. Natalello, E. Papaleo, L. De Gioia, S.M. Doglia, M. Lotti, S. Brocca, Reciprocal influence of protein domains in the cold-adapted acyl aminoacyl peptidase from *Sporosarcina psychrophila*, PLoS One 8 (2013) e56254.
[33] E. Goormaghtigh, V. Raussens, J.-M. Ruysschaert, Attenuated total reflection infrared spectroscopy of proteins and lipids in biological membranes, Biochim. Biophys. Acta Rev. Biomembr. 1422 (1999) 105–185.
[34] A. Natalello, A. Relini, A. Penco, L. Halabelian, M. Bolognesi, S.M. Doglia, S. Ricagno, Wild type beta-2 microglobulin and DE loop mutants display a common fibrillar architecture, PLoS One 10 (2015) e0122449.
[35] U.K. Laemmli, Cleavage of structural proteins during the assembly of the head of bacteriophage T4, Nature 227 (1970) 680–685.
[36] L.P. Kozlowski, IPC–isoelectric point calculator, Biol. Direct 11 (2016) 55.
[37] B. Xue, R.L. Dunbrack, R.W. Williams, A.K. Dunker, V.N. Uversky, PONDR-FIT: a meta-predictor of intrinsically disordered amino acids, Biochim. Biophys. Acta Protein Proteomics 1804 (2010) 996–1010.
[38] A.S. Holehouse, R.K. Das, J.N. Ahad, M.O.G. Richardson, R.V. Pappu, CIDER: resources to analyze sequence-ensemble relationships of intrinsically disordered proteins, Biophys. J. 112 (2017) 16–21.
[39] D.G. Isom, C.A. Castañeda, B.R. Cannon, Large shifts in pKa values of lysine residues buried inside a protein, Proc. Natl. Acad. Sci. 108 (2011) 5260–5265.
[40] M.S. Lawrence, K.J. Phillips, D.R. Liu, Supercharging proteins can impart unusual resilience, J. Am. Chem. Soc. 129 (2007) 10110–10112.
[41] C.A. Charneski, L.D. Hurst, Positively charged residues are the major determinants of ribosomal velocity, PLoS Biol. 11 (2013) e1001508.
[42] A. Barth, Infrared spectroscopy of proteins, Biochim. Biophys. Acta Bioenerg. 1767 (2007) 1073–1101.
[43] A. Natalello, D. Ami, S.M. Doglia, Fourier transform infrared spectroscopy of intrinsically disordered proteins: measurement procedures and data analyses, intrinsically disordered protein analysis, Methods Exp. Tool 1 (2012) 229–244.
[44] F. Yang, The Molecular Structure of Green Fluorescent Protein, (1997).
[45] R.M. Kramer, V.R. Shende, N. Motl, C.N. Pace, J.M. Scholtz, Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility, Biophys. J. 102 (2012) 1907–1915.
[46] C. Højgaard, C. Kofoed, R. Espersen, K.E. Johansson, M. Villa, M. Willemoës, K. Lindorff-Larsen, K. Teilum, J.R. Winther, A soluble, folded protein without charged amino acid residues, Biochemistry 55 (2016) 3949–3956.
[47] R. Halfmann, A glass menagerie of low complexity sequences, Curr. Opin. Struct. Biol. 38 (2016) 18–25.
[48] C.P. Brangwynne, P. Tompa, R.V. Pappu, Polymer physics of intracellular phase transitions, Nat. Phys. 11 (2015) 899–904.

Supplementary figures

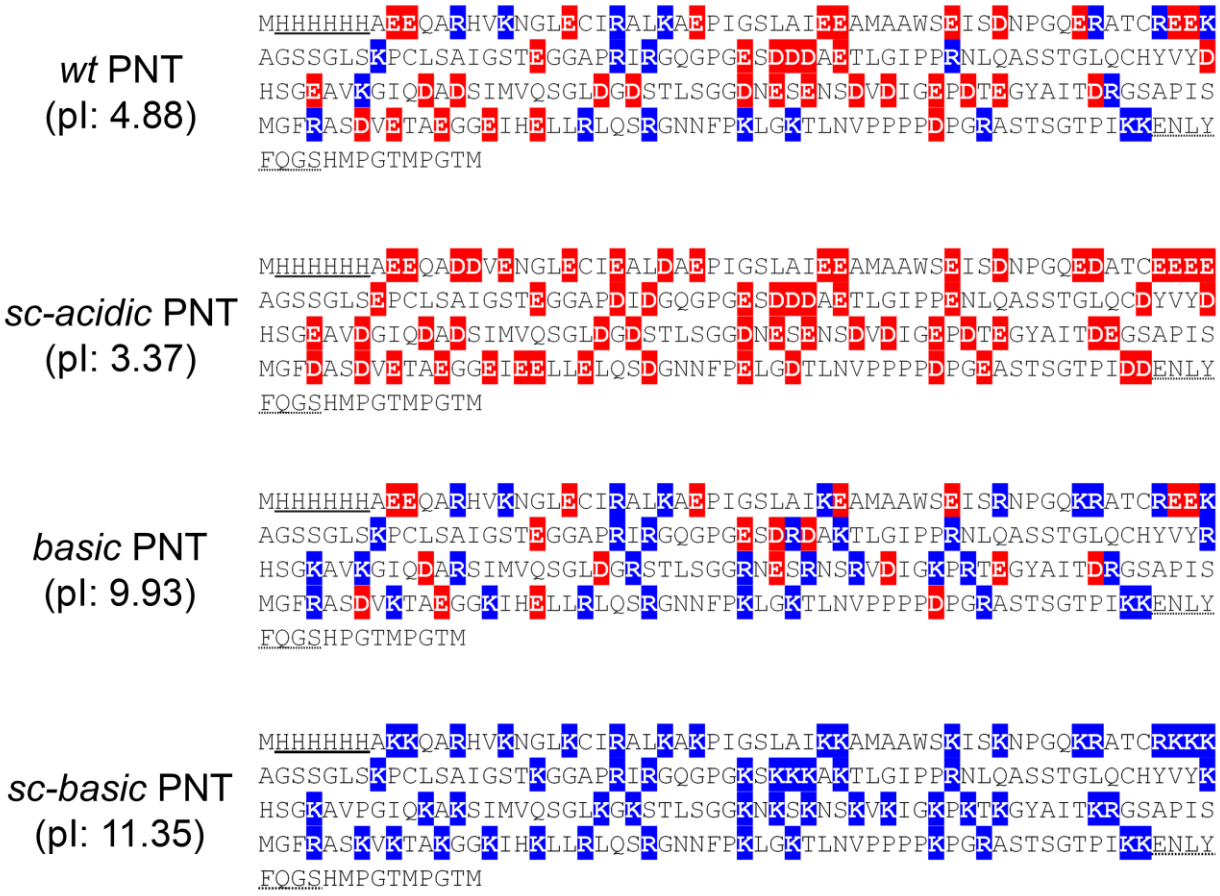


Figure S1. Sequence of PNTs variants: for each sequence, acidic and basic residues are highlighted in red and blue respectively. 6xHis tag is underlined in continuous line and TEV protease site is underlined in dotted line.

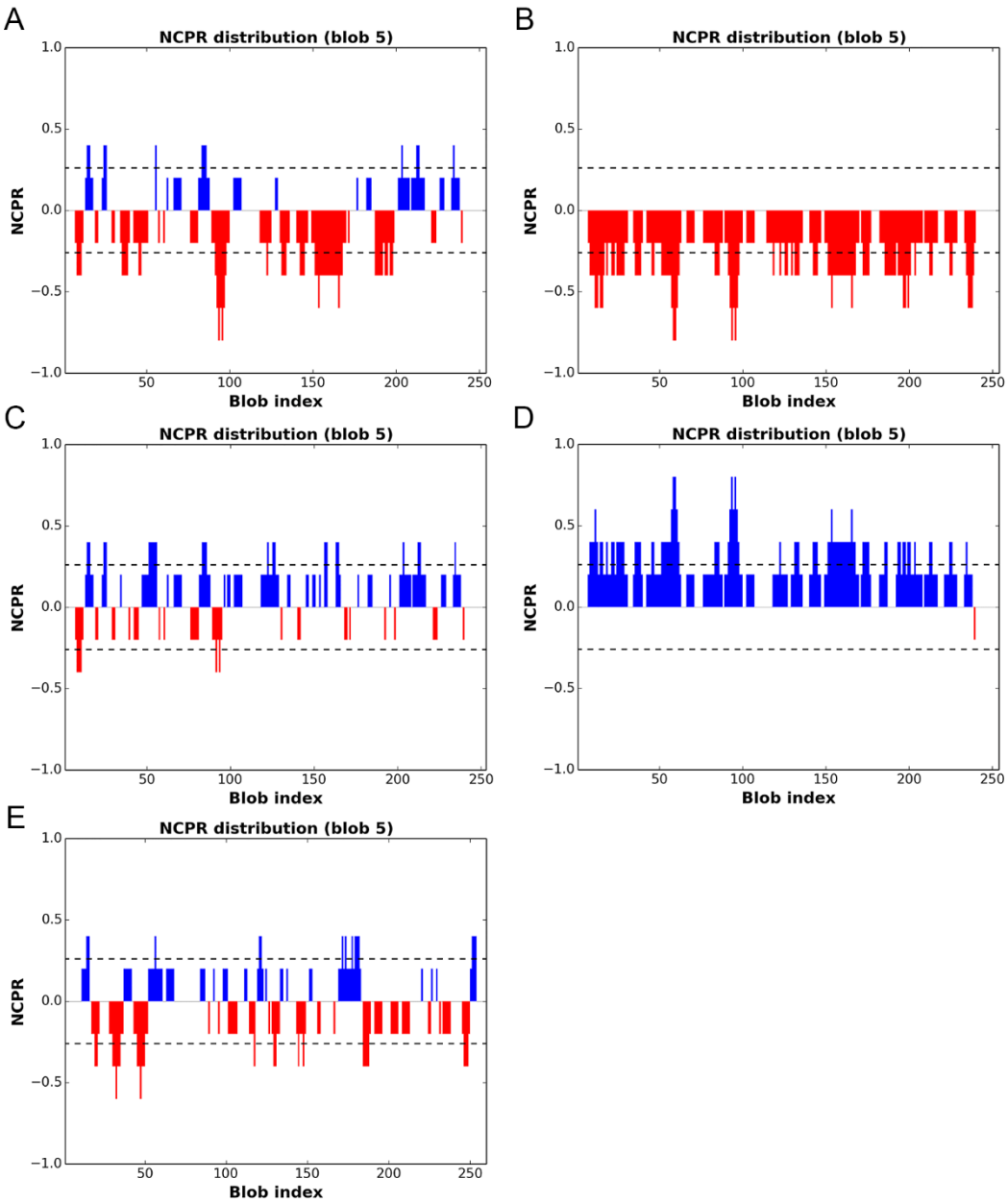


Figure S2. Linear NCPR plots. Calculations were carried out using CIDER webserver [37] with a sliding-window ("blob") of five residues. Blue and red denote positive and negatively charged residues, respectively. **A**) *wt*, **B**) *sc-acidic*, **C)** *basic*, **D)** *sc-basic* variants of PNT; **E)** GFP.