Universidad de Alcalá

COMISIÓN DE ESTUDIOS OFICIALES
DE POSGRADO Y DOCTORADO

ACTA DE EVALUACIÓN DE LA TESIS DOCTORAL

Año académico 2019/20

DOCTORANDO: **CHAVES ARQUERO, BELÉN**
D.N.I./PASAPORTE: ******5066L**

PROGRAMA DE DOCTORADO: **D424-QUÍMICA MÉDICA**
DPTO. COORDINADOR DEL PROGRAMA: **QUÍMICA ORGÁNICA Y QUÍMICA INORGÁNICA**
TITULACIÓN DE DOCTOR EN: **DOCTOR/A POR LA UNIVERSIDAD DE ALCALÁ**

En el día de hoy 21/11/19, reunido el tribunal de evaluación nombrado por la Comisión de Estudios Oficiales de Posgrado y Doctorado de la Universidad y constituido por los miembros que suscriben la presente Acta, el aspirante defendió su Tesis Doctoral, elaborada bajo la dirección de **MARIA ANGELES JIMÉNEZ LÓPEZ // JOSÉ MANUEL PÉREZ CAÑADILLAS.**

Sobre el siguiente tema: *INTRINSICALLY DISORDERED PROTEINS: STRUCTURE, DYNAMICS AND MOLECULAR RECOGNITION USING NMR SPECTROSCOPY*

Finalizada la defensa y discusión de la tesis, el tribunal acordó otorgar la CALIFICACIÓN GLOBAL[1] de (**no apto, aprobado, notable** y **sobresaliente**): _SOBRESALIENTE_

Alcalá de Henares, 21 de noviembre de 2019

EL PRESIDENTE   EL SECRETARIO   EL VOCAL

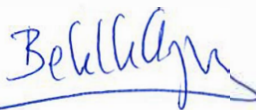Fdo.: ENCARNACIÓN MARTÍNEZ SALAS   Fdo.:DOUGLAS VINSON LAURENTS   Fdo.: FRANCISCO JOSÉ BLANCO
GUTIÉRREZ

Con fecha 25 de noviembre de 2019 la Comisión Delegada de la Comisión de Estudios Oficiales de Posgrado, a la vista de los votos emitidos de manera anónima por el tribunal que ha juzgado la tesis, resuelve:
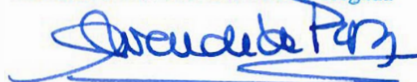
☒ Conceder la Mención de "Cum Laude"
☐ No conceder la Mención de "Cum Laude"

La Secretaria de la Comisión Delegada

FIRMA DEL ALUMNO,

Fdo.: CHAVES ARQUERO, BELÉN

---

[1] La calificación podrá ser "no apto" "aprobado" "notable" y "sobresaliente". El tribunal podrá otorgar la mención de "cum laude" si la calificación global es de sobresaliente y se emite en tal sentido el voto secreto positivo por unanimidad.

UNIVERSIDAD DE ALCALÁ, PATRIMONIO DE LA HUMANIDAD

INCIDENCIAS / OBSERVACIONES:

En aplicación del art. 14.7 del RD. 99/2011 y el art. 14 del Reglamento de Elaboración, Autorización y Defensa de la Tesis Doctoral, la Comisión Delegada de la Comisión de Estudios Oficiales de Posgrado y Doctorado, en sesión pública de fecha 25 de noviembre, procedió al escrutinio de los votos emitidos por los miembros del tribunal de la tesis defendida por **CHAVES ARQUERO, BELÉN**, el día 21 de noviembre de 2019, titulada, *INTRINSICALLY DISORDERED PROTEINS: STRUCTURE, DYNAMICS AND MOLECULAR RECOGNITION USING NMR SPECTROSCOPY* para determinar, si a la misma, se le concede la mención "cum laude", arrojando como resultado el voto favorable de todos los miembros del tribunal.

Por lo tanto, la Comisión de Estudios Oficiales de Posgrado y Doctorado **resuelve otorgar** a dicha tesis la

MENCIÓN "CUM LAUDE"

EL VICERRECTOR DE INVESTIGACIÓN Y TRANSFERENCIA
F. Javier de la Mata de la Mata
Documento fechado y firmado digitalmente

Copia por e-mail a:
Doctorando: CHAVES ARQUERO, BELÉN
Secretario del Tribunal: DOUGLAS VINSON LAURENTS
Directores de Tesis: MARIA ANGELES JIMÉNEZ LÓPEZ // JOSÉ MANUEL PÉREZ CAÑADILLAS

UNIVERSIDAD DE ALCALÁ, PATRIMONIO DE LA HUMANIDAD

Universidad
de Alcalá

DILIGENCIA DE DEPÓSITO DE TESIS.

Comprobado que el expediente académico de D./Dª _____

reúne los requisitos exigidos para la presentación de la Tesis, de acuerdo a la normativa vigente, se

procede, con fecha de hoy a registrar el depósito de la tesis en el Servicio de Estudios Oficiales de

Posgrado, con número de páginas: _____.

Alcalá de Henares a _____ de _____ de 20_____

Fdo. El Funcionario

INTRINSICALLY DISORDERED PROTEINS

STRUCTURE, DYNAMICS AND MOLECULAR RECOGNITION USING NMR SPECTROSCOPY

Universidad de Alcalá

BELÉN CHAVES ARQUERO

2019

Programa de Doctorado en Química Médica

INTRINSICALLY DISORDERED PROTEINS: STRUCTURE, DYNAMICS AND MOLECULAR RECOGNITION USING NMR SPECTROSCOPY

Tesis doctoral presentada por

BELÉN CHAVES ARQUERO

DIRECTORES

Dra. Mª Ángeles Jiménez López

Dr. José Manuel Pérez Cañadillas

TUTOR

Prof. Juan José Vaquero López

Alcalá de Henares, Septiembre 2019

Dra. M. Angeles Jiménez y Dr. José M. Pérez-Cañadillas, investigadores en plantilla del Departamento de Química Física Biológica del Instituto de Química Física Rocasolano (IQFR) del CSIC, y Prof. Juan José Vaquero López, catedrático y director del C.A.I de "Química Aplicada y Biotecnología" de la Universidad de Alcalá (UAH)

INFORMAN que la Tesis doctoral titulada "Intrinsically disordered proteins: Structure, dynamics and molecular recognition using NMR spectroscopy", presentada por Dña. Belén Chaves Arquero, y realizada bajo su dirección y tutoría reúne los requisitos científicos de originalidad y rigor metodológicos para ser defendida ante un tribunal.

Para que así conste y surta los efectos oportunos, se firma el presente informe en Madrid a 3 de septiembre de 2019
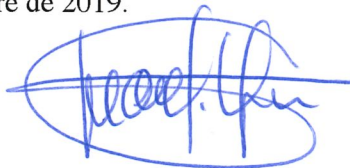
Dr. José M. Pérez Cañadillas
Codirector

Dra. M. Angeles Jiménez
Codirectora

Prof. Juan José Vaquero López
Tutor

Universidad de Alcalá

Prof. Juan José Vaquero López, Coordinador de la Comisión Académica del Programa de Doctorado en **Química Médica D424**

INFORMA que la Tesis Doctoral titulada **INTRINSICALLY DISORDERED PROTEINS: STRUCTURE, DYNAMICS AND MOLECULAR RECOGNITION USING NMR SPECTROSCOPY**, presentada por D/Dª **Belén Chaves Arquero**, bajo la dirección del / de la Dr/a. **Mª Ángeles Jiménez López y José Manuel Pérez Cañadillas, re**úne los requisitos científicos de originalidad y rigor metodológicos para ser defendida ante un tribunal. Esta Comisión ha tenido también en cuenta la evaluación positiva anual del doctorando, habiendo obtenido las correspondientes competencias establecidas en el Programa.

Para que así conste y surta los efectos oportunos, se firma el presente informe en Alcalá de Henares a 13 de Septiembre de 2019.

Fdo.: JUAN J. VAQUERO

Universidad de Alcalá

INFORME DEL DIRECTOR/A DE TESIS SOBRE "ANÁLISIS DE COINCIDENCIAS" DE LA TESIS DOCTORAL A TRAVES DEL PROGRAMA TURNITIN

D./Dª **Mª ÁNGELES JIMÉNEZ LÓPEZ** y **JOSE MANUEL PÉREZ CAÑADILLAS**, director/es de tesis, AUTORIZA/N a D/Dª **BELÉN CHAVES ARQUERO** defender la Tesis Doctoral que tiene por título **INTRINSICALLY DISORDERED PROTEINS: STRUCTURE, DYNAMICS AND MOLECULAR RECOGNITION USING NMR SPECTROSCOPY**.

El trabajo presentado ha sido analizado por la plataforma TURNITIN, arrojando un porcentaje de coincidencias del __18__ %.

Justificación de la autorización (obligatorio en caso de que el porcentaje supere el 24%):

EL 14% DE COINCIDENCIAS SE DEBE AL CAPÍTULO 2 DE LA TESIS PREVIAMENTE PUBLICADO

En Alcalá de Henares, a 13 de Septiembre de 2019

Dra. Mª Ángeles Jiménez López

Dr. José Manuel Pérez Cañadillas

Firmado: _____

Firmado: _____

Informada la Comisión Académica:

Fecha:

Firmado: Coordinador de la Comisión Académica del Programa

Prof. Juan José Vaquero López

A mi abuela

Mirar hacia atrás estos casi cuatro años solo puede hacerme valorar aun más a todas aquellas personas con las que he tenido la gran suerte de compartirlos. En primer lugar agradecer a Jose y a Mª Ángeles todo el esfuerzo empeñado en enseñarme a descubrir, a hacer ciencia y a crecer en todos lo sentidos. Vuestra distinta forma de trabajar me ha enseñado mucho de cada uno de vosotros. Gracias por brindarme esta oportunidad, esta experiencia y por inculcarme la curiosidad y las ganas de luchar por lo que se quiere conseguir. He aprendido muchísimo de vosotros.

Gracias a Marta, por ser además de un referente científico una guía de la experiencia, aquella que sabe qué decir y cómo actuar en cualquier momento. Espero seguir aprendiendo mucho de ti. Gracias al grupo de RMN, Carlos, Doug, Irene, Javi, Miguel, David, Angélica, Héctor por compartir momentos de café y pasillos, charlas necesarias y alguna que otra tarde de maqueta-foam. Gracias también a Pau Bernadó y a Nathalie Sibille por cuidarme tanto en Montpellier, fue un verdadero placer compartir esa etapa científica de mi vida con vosotros.

No me puedo olvidar de la familia rocasauria, aunque la incorporación de algunos de nosotros haya sido tardía conoceros era totalmente necesario. GRACIAS por cada momento, por cada risa, por cada anécdota, por cada plan fallido, y por todas y cada una de las comida-historia o comida-salida. Gracias a Paula, Mon, Noes, Raquel, Mompeán, Gustavo, Garavís…, a Cris por ser más que una compañera de despacho, a Sergio por compartir experimental y personalmente cada experiencia del día a día y a Isra por compartir la vida FPI del primer al casi último momento. Gracias a todos por estos años. Siempre será un placer encontrarnos, aunque el lugar tenga que estar dentro del rango de km Garavís.

Quiero agradecer a mis amigas, por estar siempre ahí, por entender cada situación y hacernos ver una vez más que la amistad no entiende de distancias. A Estefanía, por hacernos partícipes de la vida de cada una, ha sido un placer compartir esta etapa contigo. Gracias ETC & ACDC, a mis biosanos, y a la Comisión, que nunca falte la quedada de la caña, la cenas improvisadas o la organización de cualquier evento sin importar cuanto tiempo pasó desde la última vez.

A Andrés, por entender la montaña rusa de emociones que conlleva este camino, por hacerme más fácil cualquier decisión, por entender mis inquietudes y hacerlas tuyas, por querer crecer juntos en todo momento, por todo el cariño que me has dado, y por hacerme así de feliz siempre.

Por último, quiero agradecer a mi familia, a mis padres y a mi hermano. Cuando miro hacia atrás no encuentro momento en el que no hayáis estado presentes. Gracias por tanto apoyo, por vivir conmigo cada segundo, y por vuestro infinito trabajo y esfuerzo para hacer posible que cada una de mis ilusiones se vean cumplidas. Es muy bonito saber que en cada etapa, paso, decisión o destino siempre, siempre estáis ahí.
Todo lo que soy, es gracias a vosotros.

ABSTRACT

The sequence-structure-function paradigm of proteins has been revolutionized by the discovery of intrinsically disordered proteins (IDPs) or domains (IDDs), which do not have any secondary structure under physiological conditions. This absence of well-defined structure seems to be fundamental to their function. This Thesis seeks to enhance the knowledge about these proteins, focusing on the study of two systems (i) the C-terminal domain of Histone H1.0 (C-H1.0) and (ii) N-terminal domain of eukaryotic translation initiation factor eIF4G1 (eIF4G1 $_{1-250}$). The main technique used to understand the molecular basis of the biological roles of these proteins at atomic resolution has been nuclear magnetic resonance (NMR), complemented by other biophysical techniques and novel computational strategies.

C-H1.0 has a key role in regulating chromatin condensation and transcription through its DNA binding, which is modulated by phosphorylation. Structural and dynamics properties of the full-length domain, which contains three phosphorylation motifs, have been characterized in its phosphorylated and non-phosphorylated states, using a novel CON-based strategy and a minimalist approach based on model peptides.

eIF4G1 is an essential factor in translation, in post-transcriptional control and stress granules. The binding sites of (i) two RNA binding proteins (Pub1 and Pab1), (ii) RNA oligonucleotides, have been identified and self-recognition events of the three proteins have been characterized. A novel computational strategy has been used to calculate a representative ensemble being one of the first examples of an IDP structure. Analysis of protein-protein interactions individually as well as the interplay among them has been done using chemical shift and intensity mapping, paramagnetic relaxation enhancements (PREs), residual dipolar couplings (RDCs) and isotope discriminated spectroscopy. A model in which the interaction networks of Pub1-Pab1-eIF4G can progress to an oligomer forms, and eventually to condensates has been proposed. This model is relevant for physical-chemical phase-transition observed during the nucleation of stress granules. All these findings advance our understanding how structure and disorder work together will be crucial for uncovering the full extent of protein function.

RESUMEN

El descubrimiento de las proteínas desordenadas IDPs o dominios IDDs, las cuales no tienen estructura secundaria en condiciones fisiológicas, ha revolucionado el paradigma de la relación secuencia/estructura/función de las proteínas. La ausencia de estructura parece ser fundamental para su función. Esta tesis busca ampliar el conocimiento actual sobre IDPs mediante el estudio de dos sistemas: (i) el dominio C-terminal de la Histona H1.0 (C-H1.0) y (ii) el dominio N-terminal del factor de inicio de la traducción eucariota eIF4G (eIF4G1 $_{1-250}$). La principal técnica usada para entender las bases moleculares de las funciones biológicas de estas proteínas a resolución atómica ha sido la resonancia magnética nuclear (RMN). Se ha obtenido información complementaria mediante la utilización de otras técnicas biofísicas y de nuevas estrategias computacionales.

C-H1.0 tiene un papel clave en la regulación de la condensación de la cromatina y en la transcripción a través de su unión al ADN, que está modulada por fosforilaciones. Se han caracterizado las propiedades estructurales y dinámicas de este dominio, que contiene tres sitios de fosforilación, tanto en el estado sin fosforilar como en el fosforilado. Para ello se ha empleado una nueva estrategia de asignación y un enfoque minimalista basado en péptidos modelo.

eIF4G1 es un factor esencial en la traducción, en el control post-transcripcional y en los gránulos de estrés. Se han identificado los sitios de unión a (i) dos proteínas de unión a ARN (Pub1 y Pab1) , y a (ii) oligonucleótidos de ARN, y se han caracterizado la interacción intermolecular de las tres proteínas consigo mismas. Utilizando una nueva estrategia computacional se ha calculado un conjunto de estructuras representativo de eIFG1, siendo éste uno de los primeros ejemplos de estructura de IDP. Las interacciones se han analizado a partir de los cambios de desplazamientos químicos y de intensidades, los datos de relajación paramagnética PRE, los acoplamientos dipolares residuales y espectros editados por isótopos. Se ha propuesto un modelo en el que las redes de interacción entre Pab1-Pub1-eIF4G1 pueden progresar a oligómeros y eventualmente a formas condensadas. Este modelo es relevante para la transición físico-química de fase observada durante la nucleación de los gránulos de estrés. Todos estos hallazgos han avanzado el conocimiento sobre cómo estructuras ordenadas y desorden se coordinan, lo que es crucial para esclarecer por completo la función proteica.

CONTENTS

CHAPTER 2. **A CON-based NMR assignment strategy for pro-rich intrinsically disordered proteins with low signal dispersion: the C-terminal domain of histone H1.0 as a case study**

CHAPTER 3. **NMR studies on DNA recognition of the intrinsically disordered C-terminal domain of Histone H1.0**

CHAPTER 4. **Effect of phosphorylation in peptides derived from the intrinsically disordered C-terminal domain of Histone H1.0**

CHAPTER 1

INTRODUCTION

This Thesis focuses on the study of some intrinsically disordered proteins (IDPs) and domains (IDDs). Therefore this introductory chapter provides an overview of the concept of intrinsically disordered proteins. In recent years, this kind of proteins, which lack ordered three-dimensional (3D) structure, has been shown to possess a wide range of important biological functions, being remarkable their role in gene expression regulation. The classification of these IDPs is described attending to structural properties, sequence characteristics and types of binding. The most useful experimental and computational strategies applicable to characterize IDPs are also explained. Among them, nuclear magnetic resonance (NMR) is the most powerful technique to study disordered domains and proteins.

1.1. HISTORICAL PERSPECTIVES OF IDPS

One of the most basic concepts in biochemistry is that all the information required for a protein to achieve its biological functions is encoded in its amino acid sequence and that the protein can only complete these functions once it has been folded into a particular and unique structure, the native state [1].

However, this dogma has been challenged over the past 40 years due to the finding of proteins whose folded state has marginal conformational stability [2], and proteins, which contain natively unfolded regions, such as Histone H1.0 [3,4] or eukaryotic translation initiation factor eIF4G. As the years passed, it became clear that many proteins and regions hold a partly or entirely disordered structure. Indeed, a multitude of expressions has been coined to describe this class of proteins: i.e. unfolded, unstructured, disordered, denatured, natively flexible or floppy among others. The collections of names emphasize some of the characteristics of such proteins and reflects the difficulties faced by researchers trying to find an appropriate way to describe the "structure" of these proteins. Subsequently, the term **'intrinsic'** became preferred to represent the behavior of a polypeptide chain under generic physiological conditions, which is primarily defined by its amino acid sequence. Then, the discovery and characterization of such structure-less proteins, which have some residual structure place out of date the terms "unstructured" and "unfolded". Instead, the expression **'disordered'** is commonly used to describe the ensemble of structurally heterogeneous populations, with different levels of internal heterogeneity, preferred lowest energy conformations [5]. For all these reasons, the terms "**intrinsically disordered proteins**" (IDPs) and "**intrinsically disordered regions or domains"** (IDRs or IDD) represent an acceptable compromise and the expressions most currently used in the field.

At first, these disordered proteins or regions were thought to be simple connectors between structural regions. Nevertheless, it was observed that the proportion of these regions increases with the organism's complexity, which prompted a greater interest in the study of the IDP properties [6]. Early in 1990, Pontius and Berg attributed most of the activity of one protein to the disordered region [7]. Based on these findings, they proposed that intrinsically disordered domains could play important roles in promoting interactions among several partners [8].

IDPs have a distinct set of specific characteristics, which make them unable to form ordered three-dimensional structures. Hence, their functionality arises in a distinct manner to the classical structure-function view of proteins with a well-structured native state. Their different characteristics provide them functional advantages, which could be summarized in high accessibility, flexibility, a multiplicity of interactions, high specificity and low affinity [9].

Over the last 25 years, the importance of IDPs has become increasingly recognized **(Figure 1)**. Last findings clearly indicate that natural sequences have evolved to possess long IDDs, which contribute to increase the functional diversity of proteins. In fact, it has been estimated that about one third of all eukaryotes proteins are intrinsically disordered or contain at least one disordered domain [10].
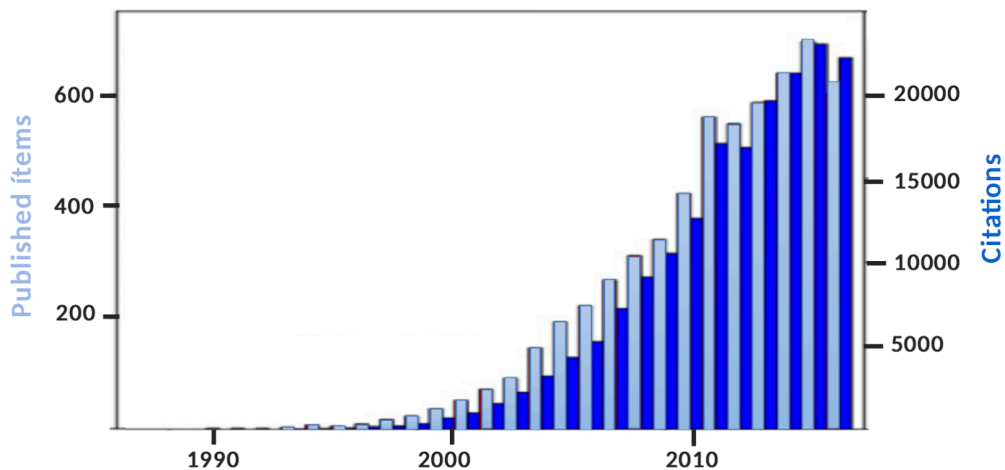
Figure 1. Number of publications in each year (light blue bars) and corresponding citations in each year (dark blue bars) related to IDPs by year, from 1991 to 2017. Publications and citations were retrieved from a search of web of Science using IDP-related terms: "(intrinsically OR natively OR inherently) and (disordered OR unstructured OR flexible) and (proteins OR proteins)".

That is, these eukaryotic proteins are generally a combination of both disordered and ordered regions, in which the disordered regions exist as dynamic ensembles of conformations. Moreover, most classical well-folded proteins still have some degree of flexibility, which is crucial for their biological activities [5].

Disordered proteins are often essential for recognition processes, signal transduction or closely linked to regulation. In all these cases, their ability to bind multiple partners, along with their

characteristic high specificity/low-affinity interactions, plays an advantageous role. As a consequence, changes in the cellular abundances or mutations in IDPs are associated with diseases, such as cancer, cardiovascular problems, diabetes and neurodegenerative disorders [11,12]. Therefore, the interactions in which IDPs are involved are attractive targets for the development a new generation of drugs [13].

In brief, the growing number of discovered IDPs, their presence in eukaryotic proteasomes, as well as the broad range of pathological and physiological functions associated with them, have fueled a great deal of enthusiasm for detailed functional and structural elucidation of IDPs. In this sense, it has to be noted that the experimental and computational methodologies, which are well established to characterize classical well-folded proteins, are not always useful for IDPs. Among the techniques used to study them (see section 1.5), the most powerful is the Nuclear Magnetic Resonance spectroscopy (NMR). Indeed, NMR is playing a leading role in developing our understanding of the physical-chemical properties of IDPs in relation to its function in health maintenance and disease progression. NMR is poised to contribute, even more, fundamental mechanistic knowledge, particularly when paired with biochemical, cellular and organism-level studies.

1.2. BIOPHYSICAL DEFINITION OF IDPS

So far this introduction has highlighted the importance of IDPs in biological function. However, the big question remains in the biophysics field about. **How can we tell whether a protein is an IDP or not?**

Proteins are linear polymers built from 20 different amino acids that are linked by peptide bonds. This simple architecture gives rise to an astronomic number of possible combinations for a protein of a particular length. For example, there are 1.26×10^{130} theoretical protein sequences of 100 amino acids. Taking into account post-translational modifications (PTMs) of some of these amino acids, the number increases even more. All of these proteins will have characteristics energetic, structural and dynamical landscapes, which are intimately correlated (**Figure 2**). This allows us to classify proteins into different categories that vary from the highly ordered to the highly disordered ones.
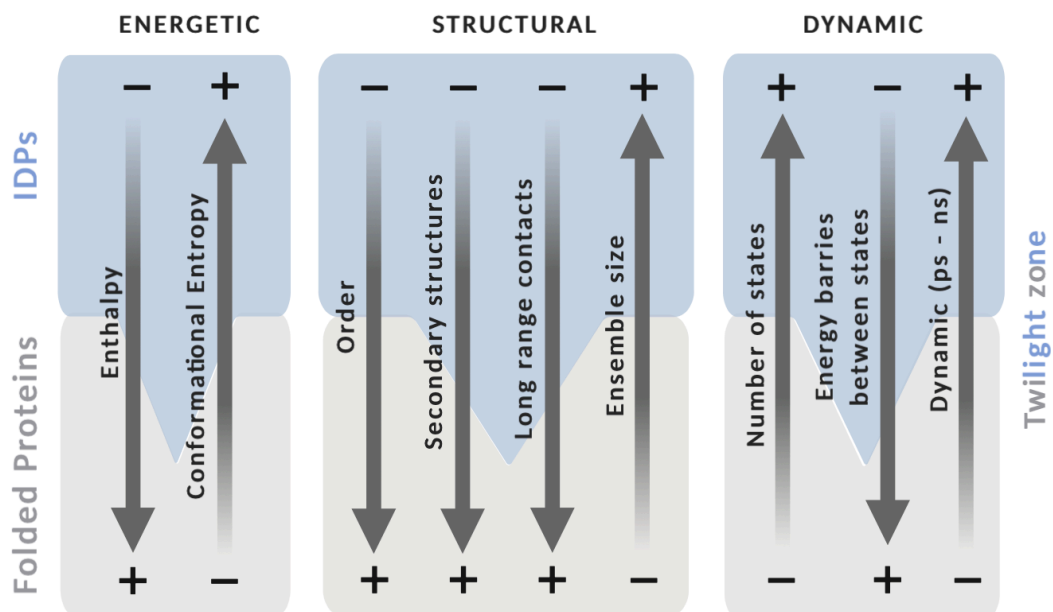

Figure 2. Correlation between characteristics energetic (enthalpy and conformational entropy), structural (order, secondary structures, long range contacts and representative ensemble size) and dynamical landscapes (number of states, energy barriers between these states and dynamic) of IDPs (blue zone) and Folded proteins (gray zone).

Two terms contribute to the protein total energy: enthalpy and entropy. Each of these terms can be further divided into other contributions (bond, angular, weak interactions, solvation, etc.), which have different "weights" for each protein. For instance, "well-folded" proteins have large **enthalpy** contributions form hydrophobic interactions (in the cores), hydrogen bonds within their secondary structure elements and other non-covalent bonds between amino acids. In contrast, IDPs lack of this favorable enthalpy terms but has higher disorder which is reflected in a high **conformational entropy (ΔS)** [14]**.** The energy landscapes between these two extremes are different. For the well-folded proteins it shows a global minimum at the major conformation and the interconversion with other states is energetically unfavorable. In contrast, the energy landscape of IDPs is proposed to be rugged, where multiple states are separated by shallow energy barriers to facilitate exchange between the states (**Figure 2**) [15,16].

In free solution, IDPs exist as an ensemble of extended and partially folded states that do not cooperatively fold into a stable structure. These IDP conformers exchange frequently among them, being each conformation functionally relevant in recognizing a specific partner for signaling or in arranging a scaffold [17]. In this regard, these IDP conformations could strategically orient specific recognition motifs such as phosphorylation sites to facilitate interaction with partner proteins, playing a critical role in reshaping the conformational landscape of IDPs. Therefore, the multiple intermediate states with ordered motifs present in the energy landscape are relevant for the implication of IDPs in gene expression regulation as we will see in section 1.4.

The concept of "conformational ensemble" acquires a full significance in IDPs, because their intrinsic structural heterogeneity can only be described by a large number of structurally different conformers. In contrast, well-folded proteins can accurately described by ensembles composed of a limited set of conformers, where variability is circumscribed only to the flexible parts (i.e. C-/N-terminal ends and long loops). Polypeptide chains can, in theory, explore a vast number of degrees of freedom. However only a subset of them are readily accessible and even less finally become part of the representative conformational ensemble of a given protein. It has been empirically determined that proteins only sample phi ϕ /psi ψ regions defined by the Ramachandran plot (**Figure 3**).

This is a way to visualize energetically allowed regions for backbone dihedral angles of amino acid residues in protein. This plot shows the empirical distribution of backbone angles gathered from experimental structures in the database. For each peptide bond, conformational ensembles of well-folded proteins show narrow distributions in the Ramachandran plot whereas IDPs usually cover all possible allowed regions reflecting there heterogeneity.
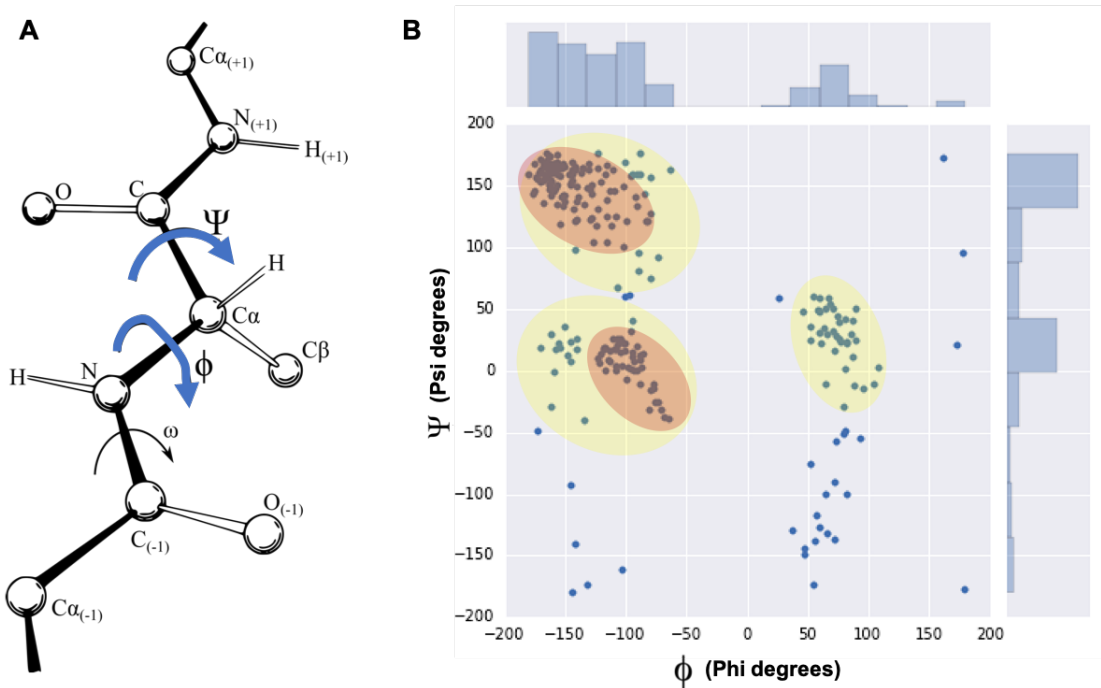
Figure 3. (A) Backbone dihedral angles ϕ and ψ in a peptide conformation are shown. **(B)** Ramachandran plot shows the allowed ϕ and ψ backbone conformational regions. The "red" regions correspond to the most favored conformations where there are no steric clashes, i.e. alpha-helical (bottom region) and beta-sheet conformations (top region). The "yellow" areas show additional allowed regions. The right region brings out an additional region which corresponds to the left-handed alpha-helix, and the rest of regions are sterically disallowed for all amino acids except glycine.

IDPs are more dynamic than well-folded proteins. This is a consequence of: (1) the low energy barriers between conformers, that allows ease jumps between them due to simple thermal fluctuations, and (2) the large number of states within their representative conformational ensembles, that permits many different pathways for interconversion.

It is clear that IDPs are defined by the correlation between characteristics energetic, structural and dynamical landscapes. But which is the best energetic/structural/dynamical parameter to decide if a protein is an IDP?. This is a difficult question to answer but efforts have been made to predict "IDP-ness" from the primary sequence (section 1.3.B), function (section 1.3.A) and structural content (section 1.3.C). These approaches are successful in predicting classic IDPs, but not proteins containing relatively high amount of structure like pre-molten globules or molten globules.

An objective parameter is required to define the degree of structureless of a particular protein. Perhaps the best tool would be conformational entropy because it provides directedly the amount of order/disorder of the system. Unfortunately, conformational entropy is not empirically measurable and often is calculated from simulations (in silico protein dynamics). However there have been several attempts to obtain empirical correlations of conformational entropy with different kinds of experimental data. A recent work shows a linear dependency of conformational entropy of Ab42 and mutant peptides with their secondary structure and the salt-bridge content [18]. Another work determines the conformational entropy changes induced upon ligand binding (TAZ1 domain of CREB binding protein) in an IDD (intrinsically disordered transactivation domain of STAT2) [19].

1.3. CLASSIFICATION OF IDPS AND IDDS

Classic well-folded proteins have unique functions due to specific interactions with partners via the binding sites. In contrast, IDPs have a more diverse and heterogenic structural landscape which increments the ability to interact, control, regulate and be controlled by partners.

Conventional protein annotation methods have difficulties to study a large pool of these IDPs and disordered proteins. One of the reasons could arise from the fact that the sequences of the disordered regions are difficult to align due to their amino acid repetitiveness. A classification scheme (**Figure 4**) is required to clarify the complex scenery of IDPs and IDDs. Such classification can be done according to (A) function, that is, how the IDP interacts with other proteins to play its functional role, (B) sequence characteristics, and (C) structural and dynamics features.
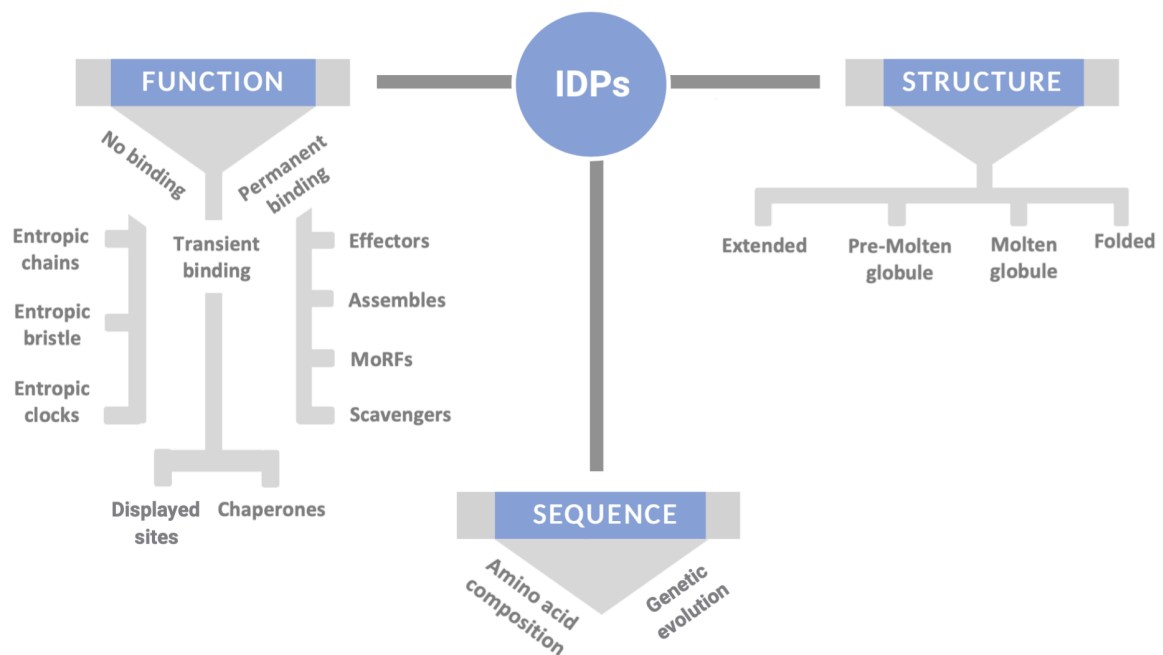
Figure 4. Classification of intrinsically disordered proteins. The proteins can be divided into three classes attending to its function, sequence and structural and dynamics features.

A. FUNCTION

Classification of IDPs according to function is difficult. Indeed, a lot of classification schemes of functional categories have been proposed based on literature analysis of more than 100 proteins containing disordered regions [20,21]. It has also to be taken into account that a single protein could contain several disordered regions, and each one might belong to a different functional class. One way of classifying them is according to whether the IDP functionality requires: (i) 'no binding' to a partner, (ii) a 'transient binding', or (iii) a 'permanent binding'.

At this point is important to recall that complex formation requires favorable free energy changes: $(\Delta G = \Delta H - T\Delta S)$. According to the definition of free energy there are two major ways for tuning the free energy of IDD to overcome the entropic cost incurred upon binding: (1) increasing the gain of enthalpy of binding ($\Delta H_{binding}$) by forming extensive interfaces or (2) by minimizing the loss of entropy. The latter can be achieved by minimizing the change in conformational entropy ($\Delta S_{conform}$) or by allowing multiple configurations for binding the interacting partner, thereby resulting in a relative gain in configurational entropy (ΔS_{config}) of the complex.

i. No binding

This category makes reference to the functions that benefit directly from the conformational disorder of the protein without ever becoming structured and without partner binding. IDPs utilize linkers and spacers of high flexibility and entropic nature to define the relative position of folded domains at each side. These segments are known as **entropic chains** and allow free movements of the domains but spatially restricted to the maximum length of the linker [22].

Among the entropic chains functions ascribed to IDPs is **entropic bristle**, which let intrinsically disordered regions sweep out a significant zone in space, so that large particles are entropically excluded whereas small molecules such as metals, cofactors, salts or water are not. An example of this mechanism is the maintenance of the axonal bore [23]. The other function is denoted **entropic clocks** because disordered domains could provide timing mechanisms, such as those observed in the ball-and-chain model for the closure of voltage-gated ion channels [24].

ii. Transient binding

The conformational flexibility of IDPs and IDDs can provide advantages over structured regions. IDPs have flexible nature domains that enclose short linear motifs (SLiMs) formed by sequences of 3-10 amino acids long. These SLiMs play an important function as **displayed sites** and facilitate post-translational modifications by enabling specific interaction with catalytic sites of modifying enzymes [25]. For example, the phosphorylation site in Histone H1.0 occurs within the disordered C-terminal domain and modification of these sites alters the function of the protein. This effect is due to the fact that the disordered region loses conformational freedom, leading to weaker and transient binding as compared to interactions with equal strength in the case of folded proteins.

Chaperones are another kind of proteins within this functional category. Intrinsically disordered regions are very common in RNA and protein chaperones, which are highly sophisticated machines that can assist the folding of proteins or nucleic acids into their biologically active states via transient binding.

The *chaperones* have disordered N- and C-terminal domains, which interact with each other, leading to the formation of oligomers. These disordered regions affect the plasticity of oligomers and also interacts with target proteins [26]. On the other hand, some non-coding RNAs that act as RNA chaperones contain *disordered tails.* These tails are nuclease-protected by interaction with an IDD of an RNA binding protein that modulate their folding and assembly [27]. Moreover, some proteins possess specific *intramolecular chaperon activities* on their C-terminal or N-terminal extensions and are required to follow their correct folding pathways [28].

Transient binding mechanism provides advantages to IDPs allowing them to structurally adapt to different partners in a fast way by increasing the number of non-specific interactions as compared to structured proteins.

iii.　Permanent binding

Another functional class of IDPs is that of the proteins, which need to interact with others to functionally work. One example is the **effectors** that interact with other proteins and modify their activity. In this case, upon binding, the disordered region often undergoes a disorder-to-order transition known as coupled folding and binding [29]. Other examples are disordered **assemblers**. Their open structure brings multiple binding partners together promoting the formation of protein complexes, that is, the disordered assemblers act as scaffolds or hubs. The complexity increases as new partners are added, and the binding sites formation are unmasked by the preceding binding stages [30].

Intrinsically disordered domains can contain **molecular recognition features** (MoRFs) giving rise to another functional class using 'permanent binding'. The MoRFs are functional regions in disordered segments composed by 10-70 amino acids, promote specific protein-protein interactions and seem to be that these features serve as initial contact points for interaction events. Both characteristics result in a large binding interface that enables multiple proteins to be bound by a single IDD. An example of a scaffold protein that uses MoRFs in the assembly process is the intrinsically disordered N-terminal domain of eukaryotic translation initiation factor eIF4G (Tif4631 in yeast). This protein assembles other RNA binding proteins, which modulate the eIF4G intramolecular interaction network. These events are proposed as critical to explain the physicochemical basis of liquid-liquid phase-transitions observed during the nucleation of stress granules (see Chapters 5, 6 and 7).

Another example of the functional class by permanent binding are **scavengers**, where the intrinsically disordered regions or proteins store and neutralize small ligands [31].

B. SEQUENCE

IDPs add new levels to classical sequence-structure relationships. After the emergence of IDPs, proteins can be crudely divided into two categories: those that can spontaneously fold into the native folded state, ordered proteins, and those that cannot, disordered proteins. This classification is possible by either (i) considering simply the amino acid sequence *per se*, or (ii) taking into account sequence evolution.

i. Amino acid composition

The intrinsic flexibility of a polymer is first governed by the nature of the bonds that connect its repeating units. All proteins, ordered and disordered, use the same types of amino acids as building blocks, which are connected via peptide bonds. However, ordered and disordered proteins differ in amino acid composition, being the percentages of each amino acid type different in both [5]. According to their differential presence in IDPs and ordered proteins, amino acid residues can be grouped into three classes: disorder-promoting residues (E, P, Q, S, R, K, M and probably D), order-promoting residues (C, W, Y, I, F, V, L, and probably N, H and T), and neutral residues (A and G) [9].

As early as 2000, Uversky et al. noticed that IDPs can be distinguished from ordered proteins attending to their average net charge and hydropathy [32]. Clearly, IDPs are enriched in proline, polar, negatively or positively charged amino acids, and reduced in large hydrophobic groups. As a consequence, they are unable to form well-organized hydrophobic cores that build up ordered domains. In addition, the sequences of IDPs contains with other two peculiarities: (1) high abundance of glutamine, asparagine and glycine-rich sequences that are implicated in amyloid formation and phase separation (section 1.4.C), and (2) high fraction of proteins with tandem repeated short segments. **Figure 5** represents these tandem repetitions in the disordered C-terminal domain of Histone H1.0. These disordered repeat regions have been shown to fall into three categories based on their functional properties [33]: (1) segments that do not have functional diversification after repeat expansion, (2) segments groups that have acquired diverse functions due to the differential location in the sequence or mutation, and (3) repeats that have gained new functions as a consequence of their expansion.
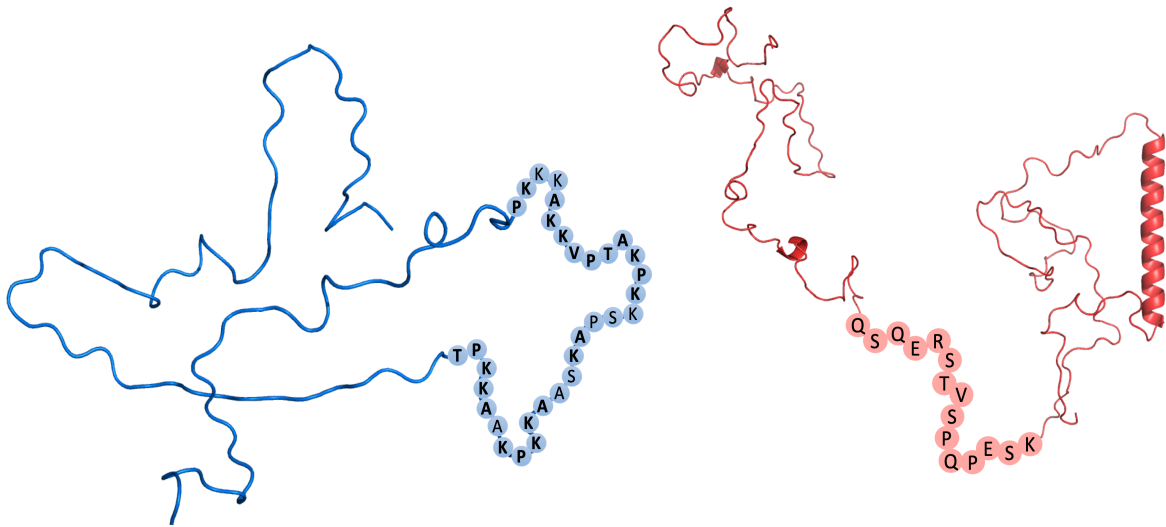
Figure 5. Schematic representation of one conformer of C-terminal domain of Histone H1.0 and N-terminal domain of eIF4G calculated with Flexible-Meccano algorithm [34] using the sequence information. The conformers represent two characteristic amino acid segments of these IDPs. Noticed that the C-terminal Histone H1.0 shows in bold the tandem repetition of short segments whereas eIF4G does not.

ii.   Genetic evolution of IDPs

Disordered regions typically evolve faster than structured domains [35,36]. In fact, the genetic instability of repetitive genomic regions found typically in these IDPs, has boosted the amount of disorder during the evolution of these particular proteins. The combination of sequence conservation of IDPs and conservation of their amino acid composition also identifies functional preferences [37]. IDPs with high residue conservation are enriched in proteins involved in nucleic acid binding and gene expression regulation. For example, IDPs involved in RNA binding and transcription regulation show high residue type conservation. Finally, neither conservation of sequence nor conservation of amino acid composition is abundant in ion binding proteins [38]. It is important to remark that increasingly complex organisms have a higher abundance of disordered in their proteomes. This suggests a further evolutionary mechanism for disordered proteins and reflects that their strong evolutionary activity and sequence-structure relation (section 1.3.C) making possible to play an important role in gene expression regulation.

C. STRUCTURAL AND DYNAMICS FEATURES

Another property of IDPs is their ability to explore different energetically similar conformations after binding to a partner or to have different percentages of residual structures. This is due to their intrinsic plasticity. The characteristics of these different states and the degree of compaction of the polypeptide chain are determined by its amino acids composition (see section 1.3.B) and their distribution. Based on their conformational behavior IDPs can be divided into fully disordered proteins, which lack of any order [4], and proteins, which are partially structured, containing both ordered and disordered domains.

Taking into account the previous IDP classification attending to function, there are IDPs that can show different structural outcomes in the bound state. These IDPs can become ordered in the bound state, and the states can be different depending on the binding partner. They can be classified in three groups: (1) IDPs, which remain disordered when they interact with their partners [39], (2) proteins that preserve significant amount of disorder in their bound state [40,41], forming the so-called fuzzy complexes (see section 1.5.A), and (3) IDPs which interact with their partners and fold concomitantly or after this interaction takes place. These structural states ranging from completely disordered to fully structured, i.e. folded domains that display no disorder loops, multidomain linked by disordered regions, compact molten globules containing extensive secondary structure and more states [38].

All these structural states are in constant exchange, since IDPs are highly dynamic and fluctuate rapidly between conformational states within the ensembles. All this complex structural landscape can be summarized and simplified in the **protein quartet model** (**Figure 6**), which proposes that protein function can arise from four conformational states (extended, pre-molten globule, molten globule and folded) and the transitions between them [42]. According to the quartet model, protein multifunctionality can depend on the degree of disorder, being different for folded state and for non-native or extended states of globular proteins. Examples of most of these cases have been found in several well-characterized proteins during the last 10 years [43]. Therefore, it is not surprising that a protein simultaneously contains regions that do not fold, folded regions and other states in between. These states might partially fold upon binding or to have partial unfolding in order to make the protein active.
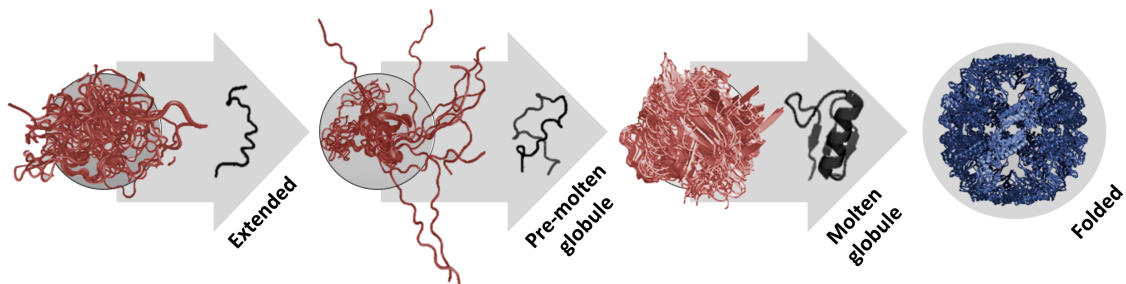
Figure 6. The protein quartet model of protein conformational states. In accordance with this model, the proteins are classified into four categories (extended, pre-molten globule, molten globule and folded), and transitions between any of these states.

In all cases, the heterogeneous view of a protein structure is constantly changing in time [44]. Globular proteins undergo a large number of interactions and energy redistributions, whereas disordered proteins majorly lack such interactions among residues. Hence, they are naturally prone to interact with others proteins or molecules, making additional contacts in order to fill this energy gap. As a result, IDPs stabilize their structures increasing the "structural heterogeneity" [45]. In fact, although folded proteins adopt well-ordered structures, they are still subjected to motions of different amplitudes occurring over a large range of timescales.

Protein dynamics is crucial for biological function, including transport processes, enzymatic catalysis and recognition events. For this reason, proteins are better represented not as a single static conformation, but as an ensemble of conformers describing protein dynamics, that play important roles in protein function [46]. For the ensemble to be complete is essential that it captures all their functional potentialities [13]. The concept of structural ensemble is more evident for IDPs, because it better represent all their characteristics. All subtypes states of IDPs play an important role in gene expression regulation, i.e. interactions where at least one of the partners is disordered in the free form, but undergoes a disorder-to-order transition upon binding.

1.4. IDPS IN GENE EXPRESSION REGULATION

Intrinsically disordered proteins or domains have been experimentally evidenced for key proteins involved in cellular processes. The database DisProt [47] provides a repository for such proteins, much of them implicated in: signaling, recognition processes and regulations activities in the cell cycle. These proteins have intrinsic disorder regions in whatever structural subtype (section 1.3.C).
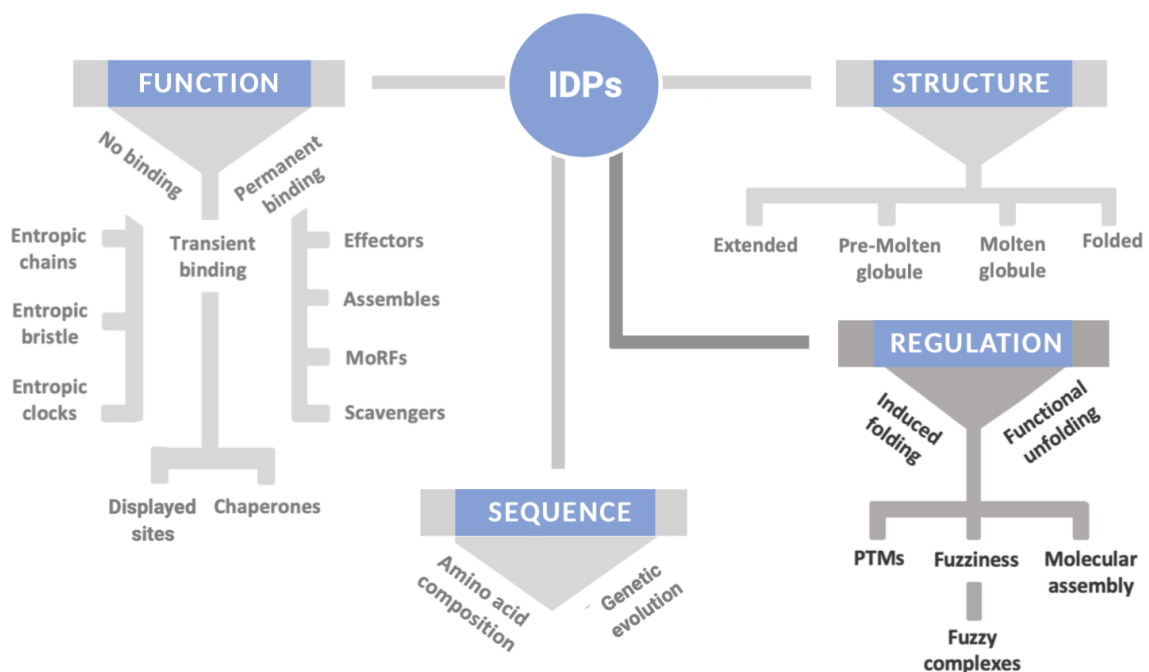
Figure 7. Classification of intrinsically disordered proteins. IDPs can be divided into three classes attending to its function, sequence, structure (section 1.3) and play an important role in gene expression regulation through three mechanisms (induced folding, functional unfolding and fuzziness).

The physical characteristics of disordered regions allow a fine level of control of cellular signaling processes (**Figure 7**). Many of these characteristics had been shown in section 1.3, and can be summarized in: the presence of small recognition elements that fold upon partner binding, accessible sites for post-translational modification, flexibility that enables them to interact with different targets, conserved sequence motifs to mediate binding interaction and the ability to bind partners with high specific but modest affinity [48–51]. The conformational characteristics of IDP ensembles are critical for the kinetics and the mechanism by which disordered interaction motifs associate with their targets.

As already mentioned, the presence of pre-formed secondary structural elements in the conformational ensemble of the IDP is important to favor binding processes [52]. However, not all IDPs undergo folding transitions to perform their biological functions. Indeed many of them remain disordered even after binding to their targets [41].

Considering all this information, we can describe three interaction mechanisms by which the IDPs will be able their function in gene expression regulation. The first mechanism is the **induced folding**, in which the protein interaction with their partner induces disorder-order transitions that promotes a regulatory response [48]. The second mechanism is the **functional unfolding**, where the protein is natively ordered and the interaction triggers its partial or complete disorder [53]. The last mechanism is also known as **fuzziness**, in which the protein maintains the intrinsic disorder upon partner binding [40,54,55]. To understand these recognition events mediated by IDPs is first necessary to comprehend the physical characteristics of disordered regions that allow the interaction to occur, and in turn trigger a specific protein function.

A. FUZZY COMPLEXES

Many IDPs can reshape their conformational ensembles upon interaction with their specific binding partner being capable of forming disordered, dynamic or "fuzzy" complexes with their partners. In this case, the fuzziness mechanisms may come in two different flavors, one of them namely "flanking fuzziness" where IDPs locally fold at the point of contact with a partner but preserve intrinsic disorder in non-contacting regions. Another type is the "random fuzziness" where IDPs remain disordered in both bound and free state [41,56,57]. The IDPs studied in this Thesis represent some examples of fuzziness mechanisms.

B. POST-TRANSLATIONAL MODIFICATIONS

The enhanced flexibility and conformational plasticity of IDPs give them readily accessible for post-translational modifications. It has been estimated that, when the PTM is taken into account, there may be as many as one million instances of peptide interaction motifs within IDRs or IDD of the human proteasome. This number underlines the central role that IDPs play in cellular regulation and signaling and support functional diversity [58].

In this regard, PTMs of an IDP provides two advantages: (1) it allows for the exposure of multiple short linear peptide motifs, which can be accessed by different enzymes to promote different signaling outputs, adding complexity to signaling pathways; and (2) the modified peptide motif can be easily accessed by effector proteins capable to recognize the PTMs in IDPs to mediate distinct outcomes.

The phosphorylation is the most common post-translational modification. Indeed the amino acid composition of the phosphorylated motifs and disordered regions (section 1.3.B) is very similar. The phosphorylation plays a major part in modulating the structural ensemble types and the interaction of disordered signaling proteins [59]. One example of phosphorylation motifs is the TPKK or TPVK sequences present in the C-terminal domain of Histone H1.0 (see Chapter 2).

C. MOLECULAR ASSEMBLY

Proteins with disordered regions can be able to interact with numerous partners. This ability allow them to be engaged in the maintenance, control and organization of complex protein-protein interaction networks, acting as **hub proteins** [60]. The ability to bind their targets through multiple sites confers on IDPs properties that facilitate the assembly of complexes, the integration of signaling pathways and also enables allosteric responses in biological signaling. In fact, the existence of multiple binding sites promotes that the energetics of the binding process is tuned by the large variations in entropy between free and bound states [14].

On the other hand, recent studies have demonstrated that the molecular assembly mechanism of IDPs can promote phase separation through liquid-liquid de-mixing to form membrane-less cytoplasmic and nuclear granules [61–64]. These granules behave like dynamic liquid droplets, rapidly exchanging components (proteins and RNA) with the cytoplasm. Granules condensate regulatory proteins and bind RNAs as an environmental response that can be regulated by high concentration or multiple PTMs (**Figure 8**).

The molecular assemblies within-granules contain a high percentage of low-complexity domains (LCDs), whose amino acid composition strongly modulate protein dynamic within the granules, than can vary in a wide range from liquid-like to solid states [65].

One key example is the intermolecular interactions among IDR of RNA-binding proteins (RBPs) [66], which are known to have a multivalent domain architecture which seems to be an important factor in phase separation (see Chapters 6 and 7). The variety of proteins and sequence biases that mediate phase separation indicates that there may be a range of underlying driving forces. These likely include electrostatic, hydrophobic, dipole-dipole, cation-pi, pi-pi, and hydrogen bonding interaction (**Figure 8**) [67,68]. It has become clear that numerous cellular organelles are formed through the process of phase separation and that IDPs play a key role in their dynamic, formation, function and regulation.
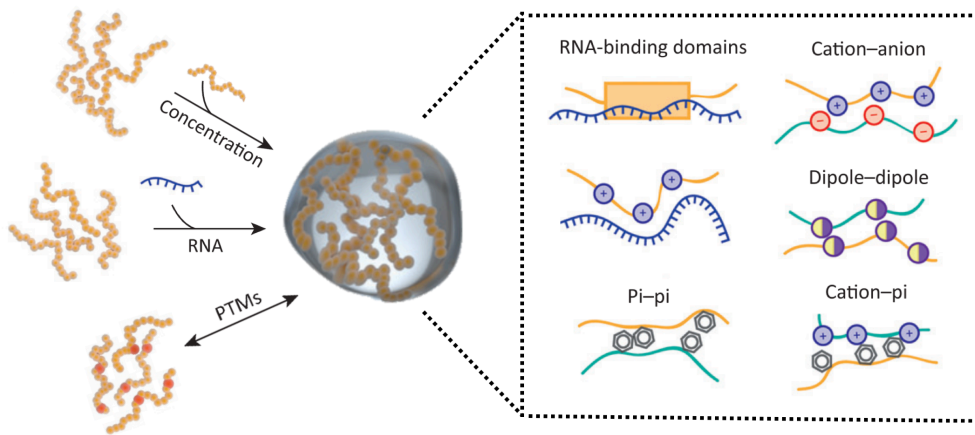
Figure 8. Scheme of interactions and regulatory mechanisms implicated in molecular assembly. On the right-hand image, is presented an overview of different kinds of contacts (RNA-binding domains, pi-pi, cation-anion, dipole-dipole and cation-pi). On the left, is found different mechanisms (high concentration, presence of RNA and post-translational modifications) that regulate the material state observed in protein phase separation.

1.5. METHODS FOR STRUCTURAL AND DYNAMICS CHARACTERIZATION OF IDPS

Characterization of the structural and dynamics properties of IDPs is of great importance to rationalize the biologically relevant roles they play. However, it poses a challenge because the experimental and computational methodologies, which are well established to characterize classical well-folded protein, are not always applicable to IDPs.

High resolution methods like X-ray crystallography and electron microscopy [69,70] are very successful to determine individual structures and large macromolecular assembles, but are incapable of catching the structural heterogeneity of IDPs. The third high resolution method is NMR, which is the most powerful technique to study IDPs structures at atomic level resolution. This is the main method used in this Thesis.
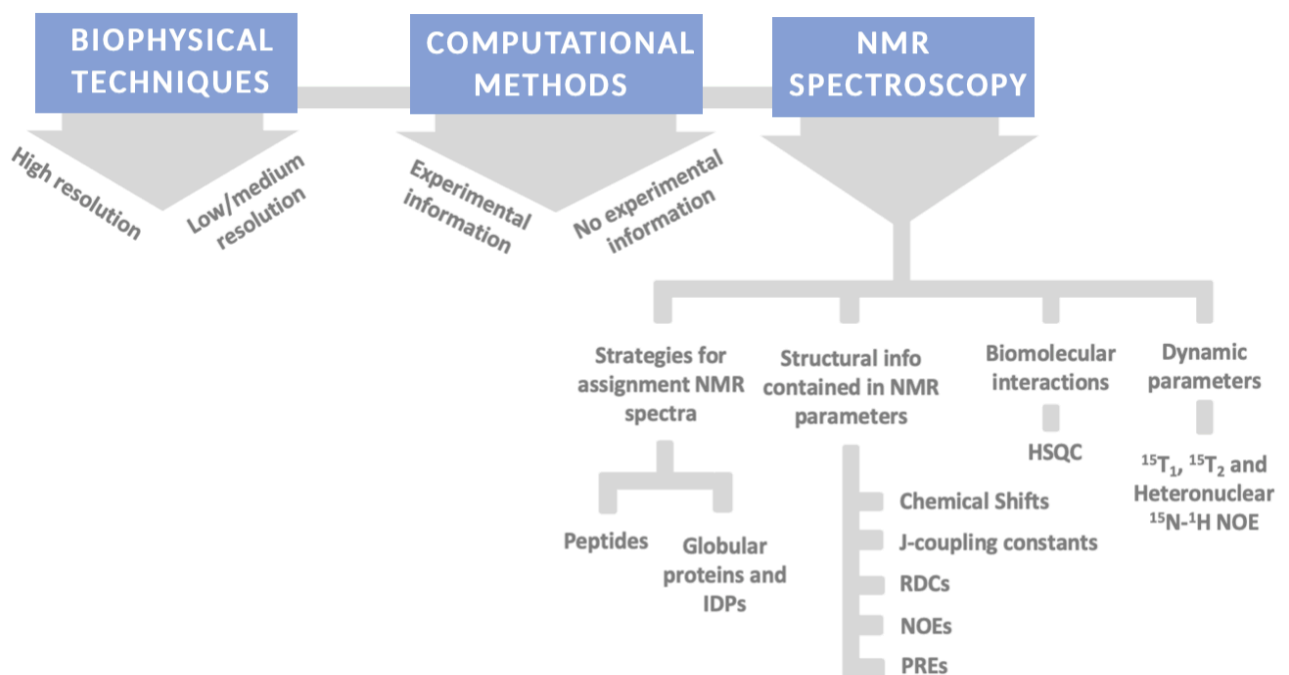
Figure 9. Scheme of methods for the structural characterization of intrinsically disordered proteins.

Recent technical advances [71] are facing the challenge posed by IDPs, and perspectives for the characterization of IDPs. In contrast to well-folded proteins, structural models of IDPs cannot be described as a single state.

Instead IDPs structures have to be represented as a collection of quickly interconverting conformers. This set is defined as "structure ensemble" and should fit the experimental parameters, that are calculated as averages across the members, despite individual structures does not.

Structural parameters are collected by different means. Best current methods for IDPs characterization include (A) biophysical techniques, (B) computational methods, and (C) NMR spectroscopy.

A. BIOPHYSICAL TECHNIQUES

Biophysical techniques provide diverse information about IDPs and their residual secondary structures. These **low/medium resolution methods** are:

- **Fluorescence** spectroscopy, including FRET (Förster resonance energy transfer) methods, which gives long-range distance between two fluorophores.
- **Infrared (IR)** spectroscopy
- **Mass spectrometry**
- **Circular dichroism (CD)**, which permits to assess the overall conformational properties of a protein or protein domain [72]
- **Confocal microscopy.** It can provide images of a complex with IDDs, i.e. oligomers or condensates.
- **Dynamic light scattering (DLS)**, which can be used to determine the size distribution profile of particles of an IDP, allowing aggregates to be identified
- **Multi-angle light scattering (MALS)**. It is used for determining both the absolute molar mass and the average size of molecules in solution.
- S**mall angle X-ray scattering (SAXS)** [73,74]. It provides data about the overall size and shape of macromolecules in solution. The samples are illuminated by a monochromatic X-ray beam, and the intensity of the scattered beam is measured at different scattering angles to get structural information [13]. In Chapter 5, we will apply this technique to analyze the interaction of an IDP and its partner.

B. COMPUTATIONAL METHODS

Algorithms used to predict secondary structures in globular proteins are useless for IDPs, due to their repetitive and low-complexity sequences (see section 1.3.B). Therefore, it has been needed to elaborate algorithms designed specifically for disorder prediction. The predictions showed that disordered regions with more than 30 consecutive residues can be found in as many as 7-30 % of prokaryotic proteins, with this number additionally increasing to 45-50% in eukaryotes [75]. These computational methods [76] can be categorized into two groups: **a**) predictors that do not use any experimental information and **b**) predictors that use experimental structural information.

The **first group (a)** includes computational methods such as PONDR models, which are based on physicochemical properties and amino acid composition [77], FoldIndex, which employs hydrophobicity and charge scores based on the prediction of disordered regions [78], DISOPRED model, which uses the sequence profiles [75,79], GlobPlot, which for each amino acid calculates the difference between its frequency of occurrence in regular secondary structures and outside of them [80], IUPred, which uses the estimation of interaction energy around a residue, and other predictors as DisEMBL [81] or NORS [82]. The **second group (b)** of methods utilizes experimental structural information to derive conformational ensembles of IDPs. These methods use NMR, SAXS data and the structural restraints derived from them to select the relevant conformations from a pool of previously generated structural ensemble. This category includes methods such as TraDES [83], Flexible-Meccano [34], ASTEROIDS [84] and ENSEMBLE [85]. All of these can generate a subset of conformers that fit available experimental data such as chemical shifts, nuclear Overhauser effect (NOE), residual dipolar coupling (RDC), paramagnetic relaxation enhancement (PRE), SAXS and hydrodynamic radius. In that way, it tests the agreement of the calculated properties to the observed experimental data.

In recent years, the combination of multiple experimental and computational studies has given insight into the structural characterization and binding mechanisms of IDPs, such as conformational transitions upon binding. However, the field of structural biology of disordered domains or proteins still needs new developments to progress in our understanding of these proteins, and to know how IDPs recognize their partner proteins with high specificity.

C. NMR SPECTROSCOPY

NMR spectroscopy allows the molecular structure of biological macromolecules to be investigated under physiological conditions. NMR active nuclei ^{1}H, ^{13}C and ^{15}N are very sensitive probes to subtle variations in the chemical environment, such as conformational changes, chemical modifications, and binding interactions (specific and non-specific). Therefore, NMR can be used with two different purposes: to obtain high-resolution structural information in peptides and proteins, and to study biomolecular interactions.

The next sections describe the NMR experiments used to obtain information about structural and dynamics features (the requirements of the protein and the assignment method), and about biomolecular interactions (using chemical shifts, intensities, residual dipolar couplings and paramagnetic relaxation enhancement) used in peptides and proteins.

i.   Strategies for assignment NMR spectra

The first step, and in many cases the bottleneck of any NMR investigation, is to assign the NMR spectra, that is, to find out the chemical shifts of every nucleus in the molecule under study. Assignment of the NMR signals of **"Peptides (10-30 residues)"** follows the strategy proposed by Wüthrich [86], which does not require the use of isotope labeling.

The experiments recorded to assign the NMR resonances following this strategy are 2D phase-sensitive two-dimensional correlated spectroscopy (COSY) and total correlated spectroscopy (TOCSY) [87], which are based on magnetization transfer between protons via scalar coupling constants ($^3J_{H-H}$) and correlates proton nuclei in the same spin system, and the Nuclear Overhauser Enhancement Spectroscopy (NOESY) [86], which have off-diagonal cross-peaks correlating protons in close proximity. These are built up by a dipole-dipole transfer of magnetization through space (Nuclear Overhauser Effect) rather than through chemical bonds. In addition, to get the ^{13}C chemical shifts the ^{1}H-^{13}C heteronuclear single quantum coherence (HSQC), which correlates ^{13}C and ^{1}H nuclei directly bound, can be recorded at natural ^{13}C abundance (1.42 %).

Assignment of **"Globular proteins and IDPs"** is not usually possible by the strategy based on 2D homonuclear spectra, and requires a combination of ^{15}N and ^{13}C labeling and 3D spectroscopy. Protein backbone assignment can be done by combining the information coming from different NMR experiments: 3D HNCO, 3D HNCA, 3D HN(CO)CA [88], 3D HN(CA)CO [89], 3D CBCANH [90] and 3D CBCA(CO)NH [91].

In all these spectra the magnetization is transferred through bonds via J-couplings, although they differ in the magnetization transfer pathways. In the 3D HNCA, 3D HNCO, 3D HN(CA)CO and 3D HN(CO)CA, magnetization starts on the proton amide and is transferred to the ^{15}N amide and then to a ^{13}C. After evolution of the ^{13}C and ^{15}N chemical shifts, the magnetization is transferred back to the proton amide for more efficient detection. This strategy is known as "out-and-back". In contrast, in the case of 3D CBCA(CO)HN and 3D CBCAHN experiments, the pathway starts on the side chain protons and travels to the amide group through several steps that include ^{13}C and ^{15}N evolution periods ("out" strategy) (**Figure 10**). It should be pointed out that "out-and-back" are longer, but more efficient strategies particularly for proteins with favorable relaxation (like IDPs). In addition, they are the only possible for perdeuterated samples.

Once assigned the backbone nuclei, a series of additional NMR experiments are acquired for the assignment of the side chain nuclei, in particular, ^{1}H-^{13}C-HSQC and two versions of 3D HCCH-TOCSY (HC(C)H-TOCSY and (H)CCH-TOCSY), which have ^{1}H-^{13}C-^{1}H or ^{13}C^{13}C^{1}H dimensions, respectively, are usually recorded.

In some cases, the NMR spectra assignment of IDPs requires the use of alternative strategies. The IDPs show unfavorable characteristics such as the exchange broadening with H_2O of ^{1}H^N signals, or that the ^{13}C$_\alpha$ and ^{13}C$_\beta$ chemical shifts tend to cluster around the random coil value for each residue type. On the other hand, their fast internal dynamics facilitates long magnetization transfers with less signal losses due to relaxation. Many different strategies have been proposed to overcome these difficulties, while taking advantage of the benefits [92–97]. One of the most important developments is the use of ^{13}C-detection methods, which has been further boosted by the new generation of NMR probes with increased ^{13}C

sensitivity. Within this area of research, we have proposed a ^{13}C-^{15}N based strategy for the C-terminal intrinsically domain of Histone H1.0 (see Chapter 2).
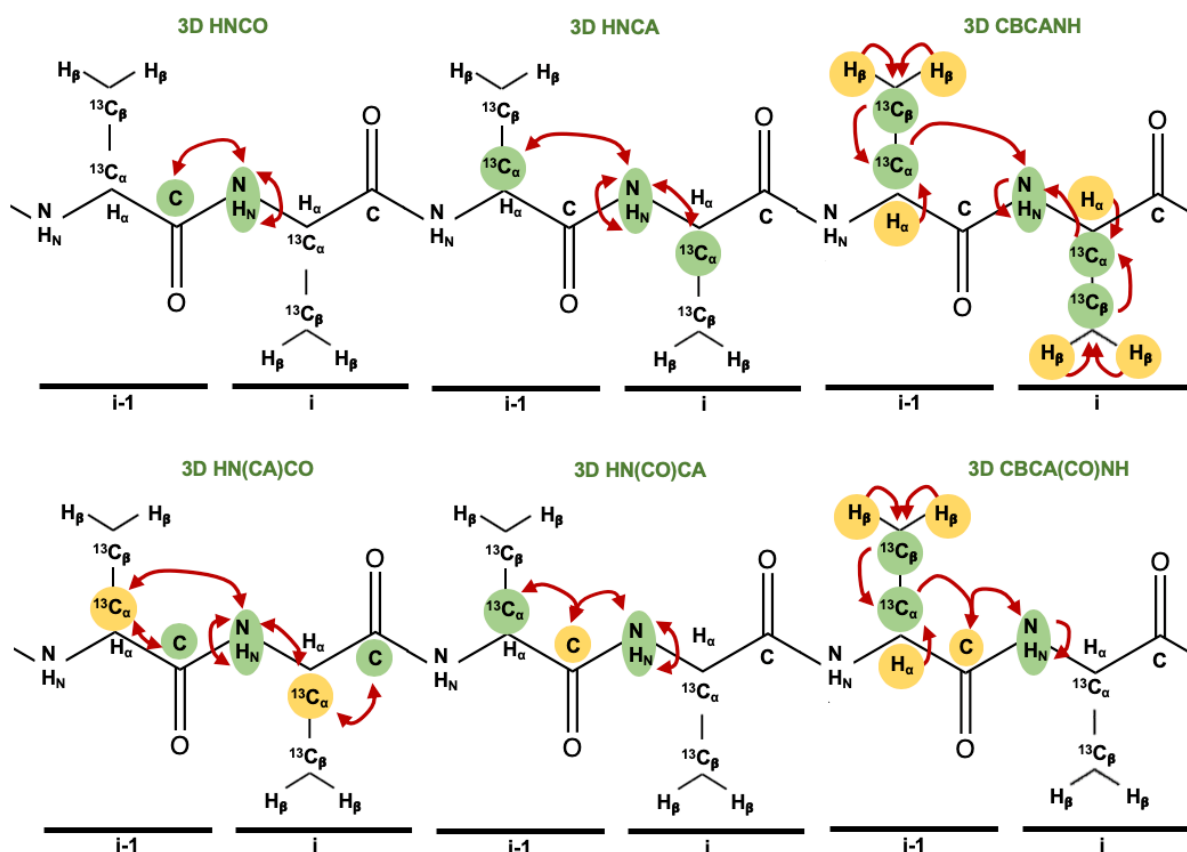
Figure 10. Magnetic transference pathways in 3D NMR experiments correlating nuclei of two sequential amino acids (i and i-1). The nuclei whose chemical shifts are recorded in the spectra are shown in green, and those nuclei that only take part in magnetic transfer pathway appear in yellow. Arrows indicate the direction of the transfers (single or double-headed for out or out-and-back experiment version, respectively).

ii.   Structural info contained in NMR parameters

Chemical shifts, with their exquisite sensitivity to environment and structure, remain the most powerful tool for identifying secondary structure not only in globular proteins, but in IDPs where secondary structure is typically transient and confined to short individual helical or extended segments [98]. The sign of the chemical shift deviations of $^1H_\alpha$, $^{13}C_\alpha$, $^{13}C_\beta$, and $^{13}C'$ is used to distinguish between helical and extended regions.

The empirical Karplus-type relationships between J-coupling constants and the bond torsion angles, **J-coupling constants** can also be used to pick up local secondary structure preference [99].

More recently, **residual dipolar couplings (RDCs)** have been shown to be sensitive to local information in disordered or unfolded proteins. The analysis of RDCs often corroborates the presence of elements of secondary structure [100]. Once the RDCs values of a given IDP are measured, new RDCs are predicted for each individual conformer of the ensemble and averaged over the entire ensemble. The prediction of RDCs relies on shape-based alignment algorithms, which are implemented in programs like Flexible-Meccano or PALES [101]. This approach has been applied to an N-terminal intrinsically domain of eIF4G (see Chapter 5). Similar to chemical shifts and scalar couplings, any deviation of experimentally measured RDCs from the calculated can be used to detect residual structures.

Medium and long-range NOEs, which are seen in two-dimensional ^{1}H-^{1}H NOESY experiment, in which cross peaks correlate spatially close protons (< 6.0 Å). At relatively low mixing times the cross peak intensities correlate linearly to the interatomic distance, more precisely with $1/r^6$, so that distance constraints can be obtained. Nevertheless, NOE information is usually scarce in disordered proteins. This likely is due to various causes: the low population of regular secondary structure elements, the high spectral overlap that makes it very difficult to reliably assign medium- or long-range NOEs and the unfavorable NOEs regime for IDPs, which make it NOEs to lay in values close to zero. Nevertheless, it has been possible to detect NOEs employing selective labeling strategies [102]. Also in proteins containing some residual structure, at it is showed the case of eIF4G in Chapter 5.

Another approach, which has proven very useful in IDPs is the use of paramagnetic centers, which strongly enhance the relaxation of nearby nuclei. Paramagnetic centers with isotropic magnetic susceptibility centers induce **paramagnetic relaxation enhancement (PRE)** dependent on the distance, as NOE, but at longer range (25-30 Å) [103]. Therefore, PRE are very useful at detecting long-range contacts in IDPs. To measure the PRE, a nitroxide spin label is attached to a single cysteine, typically engineered by mutagenesis in the protein.

In the oxidized state (paramagnetic) the dipolar interaction between the unpaired electron of this spin label and nuclear spins will increase the relaxation rate of nuclear magnetization, that is, broaden the signal. In the reduced state (diamagnetic), this PRE effect will be absent and it is used as a reference. Ratios between peaks intensities of the two states (paramagnetic and diamagnetic) can be used to calculate the PRE [104,105]. The profiles obtained with the PRE effects are useful to reveal the intricate networks of intramolecular long-range contacts between different parts of a particular protein (see Chapter 5).

3D structures of peptides and globular proteins can be determined on the basis of restraints derived from NMR parameters: distance restraints derived from NOEs, and angular restraints derived from J-coupling constants and/or chemical shifts of backbone atoms ($^{1}H_\alpha$, $^{13}C_\alpha$, $^{13}C_\beta$,$^{13}C'$). In this Thesis, structures have only been calculated for the peptides studied in Chapter 4.

iii.   Biomolecular interactions

The interaction of a protein, both ordered and IDP, with any partner affect the NMR parameters. In particular, the chemical shifts, which are very sensitive to the environment, are different for a protein in its free state and upon interaction with diverse partners. Thus, the easiest way to study the structural changes that accompany biomolecular interactions is to compare spectra acquired for the free protein and for the protein in the presence of its partner. To this end, **^{1}H-^{15}N HSQC experiment**, in which cross-peaks for all backbone amide groups (except prolines), as well as side chains groups of asparagine, arginine, glutamine and tryptophan [106] are observed, is commonly used. Usually titration experiments are followed throughout ^{1}H-^{15}N HSQC spectra, which are acquired for the free protein and at different interacting partner concentrations. Also, this experiment can be used to compare different constructs of the same protein in different conditions. The changes of intensities and positions of the spectral peaks are analyzed and quantified using Equation 1.

Series of HQSC-based experiments can be also used to derive relaxation rates. All of these parameters provide local information of the proteins. Other parameters, like PRE, can also be used to obtain long-range information of intermolecular interaction between proteins.

Equation 1.

$$\Delta\delta^{av} = \sqrt{\left((\Delta\delta_{1H})^2 + \left(\Delta\delta_{15N}/5\right)^2\right) \cdot 0.5}$$

iv.   Dynamic parameters: $^{15N}T_1$, $^{15N}T_2$ and Heteronuclear ^{15}N-^1H NOE

Protein backbone dynamics are typically studied by measuring ^{15}N-relaxation parameters: $^{15N}T_1$, $^{15N}T_2$ and Heteronuclear ^{15}N-^1H NOE. The global and local dynamic time scales can be studied by model free formalism or by spectral density mapping. In the case of IDPs, because they do not have an unique global structure, they also lack of a narrow and well-defined global correlation time. In these cases, it is advisable to study local correlation times that can be easily derived from T_1/T_2 ratios (like in Chapters 5 and 6). These motions typically occur in the nanosecond time scales.

Also the Heteronuclear ^{15}N-^1H NOE experiment [107] provides valuable information on protein dynamics in the fast time scale (picoseconds to nanoseconds). Backbone heteronuclear ^{15}N-^1H NOEs identify motions of individual N-H bond vectors that are faster than the overall tumbling of the molecules, which show a decreased NOE intensity relative to the average. Internal motions affect the rate at which an excited nucleus may sample the surrounding fluctuating fields that cause energy exchange and relaxation. Examination of the experimental data immediately reveals the distinct different motional regimes across the protein backbone. Heteronuclear NOE values close to the maximum theoretical value (which is close to 1) are indicative of rigid or ordered residues, and negative and small positive values are observed in flexible regions. In ordered proteins, the N-terminal regions and certain loops can show low or even negative heteronuclear NOE values, which indicate that these regions are more flexible than the rest of the protein. IDPs usually show an averaged low value for the heteronuclear NOE, and sometimes they can present regions with higher values, and so slightly less flexible than the rest of the IDP. Transient long range contacts can also affect relaxation and, depending on the time scale, can introduce chemical exchange contributions to $^{15N}T_2$, that reduce its value in comparison with other (like it is showing Chapter 5).

In summary, NMR is likely the best method to characterize the structure and dynamics of IDPs. NMR advances are contributing to make easier to assign NMR spectra, detect residual secondary structure or characterize interaction with other proteins. In consequence, the NMR applications developed for IDPs have shown a tremendous boost during the last years. NMR of IDPs is still a growing field where further advances are expected to happen in the near future in order to face the remaining and coming challenges in the IDPs structural biology area [108].

1.6. OBJECTIVES AND ROUTE MAP OF THIS THESIS

As it has been mentioned throughout the Introduction section, one of the great "knowledge gaps" is related to the mechanisms of interaction involving disordered zones of macromolecules (IDDs and IDPs), which play important biological roles. For instance, we have highlighted the importance of the IDPs in gene expression regulation, where they play a key role as coordinators of multicomponent complexes, and are essential to grant functionality and synchronization to the multiple cell processes. Because of this role many of these proteins are on the target for the development of pathological states that occur as a consequence of defective molecular recognition mechanisms or as misbalance in their aggregation/oligomerization properties. Nonetheless, IDPs Hubs typically contain prion-like sequences, which are important for their function, but that sometimes can evolve to toxic fibrils, that constitute an active field of structural biology studies. In contrast, the structural knowledge of IDPs in their soluble state is still on its infancy. This is due to the fact that IDDs and IDPs are difficult to characterize by most biophysical techniques, being NMR the best suitable to provide atomic details about them. Important theoretical and experimental developments in the frontiers between chemistry and biology would allow us to reveal the secrets of this key type of proteins.

In this context, the general objective of this Thesis is to contribute to advance the knowledge of the mechanisms of molecular recognition in IDPs and IDDs. To this aim, we have focused on the study of two biologically relevant systems, in which IDDs are key players: the C-terminal domain of Histone H1.0 (**Chapters 2-4**) and the N-terminal domain of eIF4G1 (**Chapters 5-7**).

A. THE C-TERMINAL IDD OF HISTONE H1.0: EFFECT OF PHOSPHORYLATION AND DNA INTERACTION

Chapters 2, 3 and 4 are dedicated to Histone 1, which is involved in the maintenance of the structure of the chromatin and in genetic regulation. Specifically, we focused on the C-terminal domain of H1.0 subtype (C-H1.0), which is intrinsically disordered and involved in DNA interaction.

C-H1.0 is classified in the transient binding category attending to its function (see section 1.3). It contains characteristic linear motifs, what makes possible the appearance of PTMs and the specific interaction with enzymes. These linear motifs represented as tandem repeated segments in C-H1.0 sequence, are significantly higher in IDPs. The C-H1.0/DNA interaction is modulated by PTMs, mainly phosphorylation's at the C-terminal domain.

The main objectives regarding to the C-H1.0 domain are:

1) To determine the structural characteristics of C-H1.0 in free state, and analyze the structural changes due to threonine's phosphorylation by using NMR. **Chapter 2** addresses this objective, which was possible thanks to a novel ^{13}C-detected CON-based strategy for NMR spectral assignment.

2) To get structural insights into the C-H1.0/DNA interaction, which is an important process in gene expression regulation, and analyze the effect of phosphorylation on DNA recognition, by using NMR. **Chapter 3** is dedicated to this objective, which could only be addressed once objective 1 was completed.

3) To study model peptides derived from C-H1.0 and explore their strength as a minimalist approach for structural characterization of IDPs. To that aim, **Chapter 4** describes the CD and NMR study of C-H1.0-derived peptides containing a single phosphorylation motif, both in non-phosphorylated and phosphorylated states. This strategy, which would be applicable to any linear PTM, would complement the studies of full-length IDDs, and could provide additional structural details.

B. THE N-TERMINAL IDD OF eIF4G1: NUCLEATION OF STRESS GRANULES AND MOLECULAR RECOGNITION

This section provides a comprehensive link between **Chapters 5, 6 and 7** presented in this thesis, which are dedicated to the study of the intrinsically disordered N-terminal domain of eIF4G1 (Tif4631 in yeast) and its interactions. This domain is a multi-tool for RNA and protein recognition with poly(A) and poly(U)-binding proteins, which will form an essential key in **(i)** the initiation of translation and **(ii)** the formation of stress granules.

(i) Translation starts with recruitment of the 43S pre-initiation complex (small ribosomal subunit 40S, eukaryotic initiator factors and the methylated initiator methionine transfer RNA or Met-tRNAiMet) to the mRNA. This process is stimulated by mRNA 3'-5' circularization. The most widely accepted mechanism for this (but not the only [109]) is the formation of an mRNA closed-loop, which is made up of Pab1, eIF4G and eIF4E. Pab1 recognized the poly(A) tail at the 3'-end of the mRNA, eIF4E factor is bound to the CAP structure of 5'-end of mRNA, and eIF4G acts as scaffold protein between Pab1 and eIF4E that makes possible the close loop formation [110,111]. Once formed, it stimulates translation initiation by promoting 43S recruitment to recognize the initiation codon prior the assemble of the whole ribosome.

(ii) The regulation of protein expression at the RNA level (under stress-induced conditions) can be characterized by the accumulation of most mRNAs into stress granules (SG) structures. Pub1 is an essential component in the formation of these membrane-less organelles (MLO) and, at the same time it also interacts with eIF4G [112,113]. Understanding the structural details in the interaction network of Pub1, Pab1 and eIF4G would likely contribute to increasing the knowledge of the SG assembly.

The main objectives regarding to the N-terminal domain of eIF4G1 are:

1) To structurally characterize the intrinsically disordered N-terminal domain of eIF4G1 (eIF4G1 $_{1-250}$) which plays an assemble function as a scaffold protein. The domain contains various MoRFs along the its sequence which promote multiple protein-protein interactions that allow it to be essential for functions (i) and (ii). The unknown structural characteristics of eIF4G1 $_{1-250}$ will be study in **Chapter 5** by implementing a novel strategy that use knowledge-

based data about the dominant interactions in IDD to generate ensembles that faithfully reproduce experimental NMR data. As minimum size structural ensemble has been generated, which explains some of the properties of eIF4G1 $_{1-250}$ and might represent a pioneering study in the IDD structural field.

2) To determine the intermolecular interactions by self-recognition of the three proteins eIF4G1 $_{1-250}$, poly(U) and poly(A)-binding proteins (Pub1 and Pab1). These RNA binding proteins have a similar structural architecture with 3 and 4 RRM domains and are described as components of ribonucleoprotein condensates because may give rise to form oligomers. This fact could be useful for the nucleation of stress granules where they play an essential role. Therefore, **Chapter 6** will focus on the understanding that the key role of the self-recognition interaction of these three proteins plays in (i), (ii), and in their interactions with other proteins.

3) To characterize the interaction network between RNA, Pub1, Pab1 and eIF4G1 $_{1-250}$. This main objective is addressed in **Chapter 7**. In this chapter, the binding sites of the RNA binding proteins in eIF4G1 $_{1-250}$ have been defined by NMR allowing the identification of new binding sites that could provide new mechanistic insights into the cap-independent translation initiation process. Peptide models will be used to corroborate the studies with the complete proteins. The combined effect of the three proteins will be studied by NMR and fluorescent confocal microscopy in order to construct a mechanistic model about the possible impact of multivalent interactions between the proteins in the nucleation of SG.

1.7. REFERENCES

1. Anfinsen CB. Principles that govern the folding of protein chains. Science (80). 1973;181(4096):223–30.

2. Pace CN, Hermans J. The stability of globular protein. Crit Rev Biochem Mol Biol. 1975;3(1):1–43.

3. Lilley DMJ, Howarth OW, Clark VM. and J. F. PARDON and B. M. RICHARDS. 1976;62(1):1–4.

4. Chaves-Arquero B, Pantoja-Uceda D, Roque A, Ponte I, Suau P, Jiménez MA. A CON-based NMR assignment strategy for pro-rich intrinsically disordered proteins with low signal dispersion: the C-terminal domain of histone H1.0 as a case study. J Biomol NMR. 2018;72(3–4):139–48. Available from: http://dx.doi.org/10.1007/s10858-018-0213-2

5. Dunker AK, Babu MM, Barbar E, Blackledge M, Bondos SE, Dosztányi Z, et al. What's in a name? Why these proteins are intrinsically disordered. Intrinsically Disord Proteins. 2013;1(1):e24157.

6. Yu JF, Cao Z, Yang Y, Wang CL, Su ZD, Zhao YW, et al. Natural protein sequences are more intrinsically disordered than random sequences. Cell Mol Life Sci. 2016;73(15):2949–57.

7. Pontius BW, Berg P. Renaturation of complementary DNA strands mediated by purified mammalian heterogeneous nuclear ribonucleoprotein A1 protein: implications for a mechanism for rapid molecular assembly. Proc Natl Acad Sci. 1990;87(21):8403–7.

8. Pontius BW. Close encounters: why unstructured, polymeric domains can increase rates of specific macromolecular association. Trends Biochem Sci. 1993;18(5):181–6.

9. Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN, Dunker AK. Intrinsic disorder and functional proteomics. Biophys J. 2007;92(5):1439–56. Available from: http://dx.doi.org/10.1529/biophysj.106.094045

10. Xue B, Dunker AK, Uversky VN. Orderly order in protein intrinsic disorder distribution: Disorder in 3500 proteomes from viruses and the three domains of life. J Biomol Struct Dyn. 2012;30(2):137–49.

11. Uversky VN, Oldfield CJ, Dunker AK. Intrinsically Disordered Proteins in Human Diseases: Introducing the D 2 Concept . Annu Rev Biophys. 2008;37(1):215–46.

12. Babu MM, van der Lee R, de Groot NS, Gsponer J. Intrinsically disordered proteins: Regulation and disease. Curr Opin Struct Biol. 2011;21(3):432–40.

13. Rezaei-Ghaleh N, Blackledge M, Zweckstetter M. Intrinsically Disordered Proteins: From Sequence and Conformational Properties toward Drug Discovery. ChemBioChem. 2012;13(7):930–50.

14. Flock T, Weatheritt RJ, Latysheva NS, Babu MM. Controlling entropy to tune the functions of intrinsically disordered regions. Curr Opin Struct Biol. 2014;26(1):62–72. Available from: http://dx.doi.org/10.1016/j.sbi.2014.05.007

15. Chebaro Y, Ballard AJ, Chakraborty D, Wales DJ. Intrinsically Disordered Energy Landscapes. Sci Rep. 2015 May 22;5:10386. Available from: https://doi.org/10.1038/srep10386

16. Frauenfelder H, Sligar SG, Wolynes PG. The energy landscapes and motions of proteins. Science (80). 1991 Dec 13;254(5038):1598 LP – 1603. Available from: http://science.sciencemag.org/content /254/ 5038/ 1598.abstract

17. Wei G, Xi W, Nussinov R, Ma B. Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? the Diverse Functional Roles of Conformational Ensembles in the Cell. Chem Rev. 2016;116 (11):6516–51.

18. Chong SH, Ham S. Conformational entropy of intrinsically disordered protein. J Phys Chem B. 2013;117(18):5503–9.

19. Lindström I, Dogan J. Dynamics, Conformational Entropy, and Frustration in Protein-Protein Interactions Involving an Intrinsically Disordered Protein Domain. ACS Chem Biol. 2018;13(5):1218–27.

20. van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, et al. Classification of Intrinsically Disordered Regions and Proteins. Chem Rev. 2014;114(13):6589–631.

21. Uversky VN. The most important thing is the tail: Multitudinous functionalities of intrinsically disordered protein termini. FEBS Lett. 2013;587(13):1891–901.

22. Daughdrill GW, Narayanaswami P, Gilmore SH, Belczyk A, Brown CJ. Dynamic behavior of an intrinsically unstructured linker domain is conserved in the face of negligible amino acid sequence conservation. J Mol Evol. 2007;65(3):277–88.

23. Brown HG, Hoh JH. Entropic exclusion by neurofilament sidearms: A mechanism for maintaining interfilament spacing. Biochemistry. 1997;36(49):15035–40.

24. Hoshi T, Zagotta WN, Aldrich RW. Biophysical and Molecular Mechanisms of Shaker Potassium Channel Inactivation Toshinori Hoshi; William N. Zagotta; Richard W. Aldrich. Science (80). 1990;250(4980):533–8.

25. Galea CA, Wang Y, Sivakolundu SG, Kriwacki RW. Regulation of cell division by intrinsically unstructured proteins: Intrinsic flexibility, modularity, and signaling conduits. Biochemistry. 2008;47(29):7598–609.

26. V. Sudnitsyna M, V. Mymrikov E, S. Seit-Nebi A, B. Gusev N. The Role of Intrinsically Disordered Regions in the Structure and Functioning of Small Heat Shock Proteins. Curr Protein Pept Sci. 2012;13(1):76–85.

27. Kucera NJ, Hodsdon ME, Wolin SL. An intrinsically disordered C terminus allows the La protein to assist the biogenesis of diverse noncoding RNA precursors. Proc Natl Acad Sci. 2011;108(4):1308–13.

28. Chen YJ, Inouye M. The intramolecular chaperone-mediated protein folding. Curr Opin Struct Biol. 2008;18(6):765–70.

29. Wright PE, Dyson HJ. Linking Folding and Binding. 2009;19(1):31–8.

30. Xue B, Romero PR, Noutsou M, Maurice MM, Stefan GD, Jr AMW, et al. Stochastic machines as a colocalization mechanism for scaffold protein function. 2013;587(11):1587–91.

31.   Holt C. Unfolded phosphopolypeptides enable soft and hard tissues to coexist in the same organism with relative ease. Curr Opin Struct Biol. 2013;23(3):420–5. Available from: http://dx.doi.org/10.1016 /j.sbi.2013.02.010

32.   Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins Struct Funct Genet. 2000;41(3):415–27.

33.   Tompa P. Intrinsically unstructured proteins evolve by repeat expansion. BioEssays. 2003;25(9):847–55.

34.   Ozenne V, Bauer F, Salmon L, Huang JR, Jensen MR, Segard S, et al. Flexible-meccano: A tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. Bioinformatics. 2012;28(11):1463–70.

35.   Nilsson J, Grahn M, Wright APH. Proteome-wide evidence for enhanced positive Darwinian selection within intrinsically disordered regions in proteins. Genome Biol. 2011;12(7):R65. Available from: http:// genomebiology.com/2011/12/7/R65

36.   Brown CJ, Johnson AK, Dunker AK, Daughdrill GW. Evolution and disorder. Curr Opin Struct Biol . 2011;21(3):441–6. Available from: http://dx.doi.org/10.1016/j.sbi.2011.02.005

37.   Moesa HA, Wakabayashi S, Nakai K, Patil A. Chemical composition is maintained in poorly conserved intrinsically disordered regions and suggests a means for their classification. Mol Biosyst. 2012; 8(12):3262–73.

38.   Van Der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, et al. Classification of intrinsically disordered regions and proteins. Vol. 114, Chemical Reviews. 2014. p. 6589–631.

39.   Uversky VN, Dunker AK. The case for intrinsically disordered proteins playing contributory roles in molecular recognition without a stable 3D structure. F1000 Biol Rep. 2014;5(January):1–12.

40.   Fuxreiter M, Tompa P. Fuzzy complexes: A more stochastic view of protein function. Adv Exp Med Biol. 2012;725:1–14.

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

42.   N. UV. Natively unfolded proteins: A point where biology waits for physics. Eur J Biochem . 2002;12(May 2001):2–12. Available from: http://onlinelibrary.wiley.com/doi/10.1046/j.0014-2956.2001.02649.x/ full%5Cnfile

43.   Yang S, Blachowicz L, Makowski L, Roux B. Multidomain assembled states of Hck tyrosine kinase in solution. Proc Natl Acad Sci. 2010;107(36):15757–62.

44.   Uversky VN. A decade and a half of protein intrinsic disorder: Biology still waits for physics. Protein Sci. 2013;22(6):693–724.

45.   Upadhyay A. Structure of proteins: Evolution with unsolved mysteries. Prog Biophys Mol Biol . 2019;(xxxx). Available from: https://doi.org/10.1016/j.pbiomolbio.2019.04.007

46.   Baldwin AJ, Kay LE. NMR spectroscopy brings invisible protein states into focus. Nat Chem Biol . 2009;5(11):808–14. Available from: http://dx.doi.org/10.1038/nchembio.238

47.   Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, et al. DisProt: The database of disordered proteins. Nucleic Acids Res. 2007;35(SUPPL. 1):786–93.

48.   Wright PE, Dyson HJ. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. J Mol Biol. 1999;293(2):321–31.

49.   Dunker AK, Brown CJ, Obradovic Z. Identification and functions of usefully disordered proteins. Adv Protein Chem. 2002;62:25–49.

50.   Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. Nat Rev Mol Cell Biol. 2015;16(1):18–29. Available from: http:// www.ncbi.nlm.nih.gov/pubmed/25531225Ahttp:// www. pubmedcentral.nih.gov /articlerender.fcgi?artid=PMC4405151

51.   Gsponer J, Madan Babu M. The rules of disorder or why disorder rules. Prog Biophys Mol Biol. 2009;99(2–3):94–103.

52.   Fuxreiter M, Simon I, Friedrich P, Tompa P. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. J Mol Biol. 2004;338(5):1015–26.

53.   Kurzbach D, Schwarz TC, Platzer G, Höfler S, Hinderberger D, Konrat R. Compensatory adaptations of structural dynamics in an intrinsically disordered protein complex. Angew Chemie - Int Ed. 2014;53(15):3840–3.

54.   Fuxreiter M. Towards a Stochastic Paradigm: From Fuzzy Ensembles to Cellular Functions. Molecules. 2018;23(11):3008.

55.   Sharma R, Demény M, Ambrus V, Király SB, Kurtán T, Gatti-Lafranconi P, et al. Specific and Fuzzy Interactions Cooperate in Modulating Protein Half-Life. J Mol Biol. 2019;431(8):1700–7. Available from: https://doi.org /10.1016/j.jmb.2019.02.006

56.   Hazy E, Tompa P. Limitations of induced folding in molecular recognition by intrinsically disordered proteins. ChemPhysChem. 2009;10(9–10):1415–9.

57.   Arbesú M, Iruela G, Fuentes H, Teixeira JMC, Pons M. Intramolecular Fuzzy Interactions Involving Intrinsically Disordered Domains. Front Mol Biosci. 2018;5(April):1–7.

58.   Tompa P, Davey NE, Gibson TJ, Babu MM. A Million peptide motifs for the molecular biologist. Mol Cell . 2014;55(2):161–9. Available from: http://dx.doi.org/10.1016/j.molcel.2014.05.032

59.   Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, et al. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res. 2004;32(3):1037–49.

60.   Kim PM, Sboner A, Xia Y, Gerstein M. The role of disorder in interaction networks: A structural analysis. Mol Syst Biol. 2008;4(May 2014):179.

61.   Kedersha N, Ivanov P, Anderson P. Stress granules and cell signaling: More than just a passing phase? Trends Biochem Sci. 2013;38(10):494–506. Available from: http://dx.doi.org/10.1016/j.tibs. 2013.07.004

62.   Li P, Banjade S, Cheng HC, Kim S, Chen B, Guo L, et al. Phase transitions in the assembly of multivalent signalling proteins. Nature. 2012;483(7389):336–40. Available from: http://dx.doi.org/10.1038/nature 10879

63.   Weber SC, Brangwynne CP. Getting RNA and protein in phase. Cell . 2012;149(6):1188–91. Available from: http://dx.doi.org/10.1016/j.cell.2012.05.022

64.   Boeynaems S, Alberti S, Fawzi NL, Mittag T, Polymenidou M, Rousseau F, et al. Protein Phase Separation: A New Phase in Cell Biology. Trends Cell Biol. 2018;28(6):420–35. Available from: http://dx.doi.org /10.1016/j.tcb.2018.02.004

65.   Uversky VN. Protein intrinsic disorder-based liquid–liquid phase transitions in biological systems: Complex coacervates and membrane-less organelles. Adv Colloid Interface Sci . 2017;239:97–114. Available from: http://dx.doi.org/10.1016/j.cis.2016.05.012

66.   Varadi M, Zsolyomi F, Guharoy M, Tompa P, Levy YK. Functional advantages of conserved intrinsic disorder in RNA-binding proteins. PLoS One. 2015;10(10):1–16.

67.   Brangwynne CP, Tompa P, Pappu R V. Polymer physics of intracellular phase transitions. Nat Phys. 2015;11(11):899–904.

68.   Vernon RM, Chong PA, Tsang B, Kim TH, Bah A, Farber P, et al. Pi-Pi contacts are an overlooked protein feature relevant to phase separation. Elife. 2018;7:1–48.

69.   Bonomi M, Hanot S, Greenberg CH, Sali A, Nilges M, Vendruscolo M, et al. Bayesian Weighing of Electron Cryo-Microscopy Data for Integrative Structural Modeling. Structure. 2019;27(1):175-188.e6. Available from: https://doi.org/10.1016/j.str.2018.09.011

70.   Bonomi M, Pellarin R, Vendruscolo M. Simultaneous Determination of Protein Structure and Dynamics Using Cryo-Electron Microscopy. Biophys J. 2018;114(7):1604–13. Available from: https://doi.org/10.1016/j.bpj. 2018.02.028

71.   Eliezer D. Biophysical characterization of intrinsically disordered proteins. Curr Opin Struct Biol. 2009;19(1):23–30.

72.   Chemes B, Alonso LG, Noval G, Prat-gay G De. Intrinsically Disordered Protein Analysis. 2012;896(Cd):387–404. Available from: http://link.springer.com/10.1007/978-1-4614-3704-8

73.   Kachala M, Valentini E, Svergun DI. Application of SAXS for the Structural Characterization of IDPs. In: Felli IC, Pierattelli R, editors. Intrinsically Disordered Proteins Studied by NMR Spectroscopy . Cham: Springer International Publishing; 2015. p. 261–89. Available from: https://doi.org/10.1007/978-3-319-20164-1_8

74.   Bernadó P, Svergun DI. Analysis of Intrinsically Disordered Proteins by Small-Angle X-ray Scattering. In: Uversky VN, Dunker AK, editors. Intrinsically Disordered Protein Analysis: Volume 2, Methods and Experimental Tools . New York, NY: Springer New York; 2012. p. 107–22.

75. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol. 2004;337(3):635–45. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15019783

76. Bhattacharya S, Lin X. Recent Advances in Computational Protocols Addressing Intrinsically Disordered Proteins. Biomolecules. 2019;9(4).

77. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein in intrinsic disorder. BMC Bioinformatics. 2006;7:1–17.

78. Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, et al. FoldIndex©: A simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics. 2005;21 (16):3435–8.

79. Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT. The DISOPRED server for the prediction of protein disorder. Bioinformatics. 2004;20(13):2138–9.

80. Linding R, Russell RB, Neduva V, Gibson TJ. GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Res. 2003;31(13):3701–8. Available from: http://www.ncbi.nlm.nih.gov /pubmed/ 12824398%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC169197

81. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: Implications for structural proteomics. Structure. 2003;11(11):1453–9.

82. Liu J, Rost B. NORSp: Predictions of long regions without regular secondary structure. Nucleic Acids Res. 2003;31(13):3833–5.

83. Feldman HJ, Hogue CWV. Probabilistic sampling of protein conformations: New hope for brute force? Proteins Struct Funct Genet. 2002;46(1):8–23.

84. Blackledge M, Jensen MR, Salmon L, Meier S, Nodet G, Ozenne V. Quantitative Description of Backbone Conformational Sampling of Unfolded Proteins at Amino Acid Resolution from NMR Residual Dipolar Couplings. J Am Chem Soc. 2009;131(49):17908–18.

85. Krzeminski M, Marsh JA, Neale C, Choy WY, Forman-Kay JD. Characterization of disordered proteins with ENSEMBLE. Bioinformatics. 2013;29(3):398–9.

86. Kumar A, Ernst RR, Wuthrich K. A two-dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton-proton cross-relaxation networks in biological macromolecules Biochemical and biophysical research communications. 1980.

87. Braunschweiler L, Ernst RR. Coherence transfer by isotropic mixing: Application to proton correlation spectroscopy. J Magn Reson. 1983;53(3):521–8.

88. Grzesiek S, Bax A. Improved 3D triple-resonance NMR techniques applied to a 31 kDa protein. J Magn Reson. 1992;96(2):432–40.

89. Clubb RT, Thanabal V, Wagner G. A constant-time three-dimensional triple-resonance pulse scheme to correlate intraresidue 1HN, 15N, and 13C' chemical shifts in 15N13C-labelled proteins. J Magn Reson. 1992;97(1):213–7.

90. Grzesiek S, Bax A. An efficient experiment for sequential backbone assignment of medium-sized isotopically enriched proteins. J Magn Reson. 1992;99(1):201–7.

91. Grzesiekt S, Bax A. Correlating Backbone Amide and Side Chain Resonances in Larger Proteins by Multiple Relayed Triple Resonance NMR. J Am Chem Soc. 1992;114(16):6291–3.

92. Mantylahti S, Hellman M, Permi P. Extension of the HA-detection based approach: (HCA)CON(CA)H and (HCA)NCO(CA)H experiments for the main-chain assignment of intrinsically disordered proteins. J Biomol Nmr. 2011/01/25. 2011;49(2):99–109. Available from: http://www.ncbi.nlm.nih.gov/pubmed /21259120

93. Bermel W, Bertini I, Felli IC, Gonnelli L, Kozminski W, Piai A, et al. Speeding up sequence specific assignment of IDPs. J Biomol Nmr. 2012/06/12. 2012;53(4):293–301. Available from: http://www.ncbi.nlm.nih.gov /pubmed/22684679

94. Nováček J, Zawadzka-Kazimierczuk A, Papoušková V, Žídek L, Šanderová H, Krásný L, et al. 5D 13C-Detected experiments for backbone assignment of unstructured proteins with a very low signal dispersion. J Biomol NMR. 2011;50(1):1–11.

95. Novacek J, Haba NY, Chill JH, Zidek L, Sklenar V. 4D non-uniformly sampled HCBCACON and (1)J(NCalpha)-selective HCBCANCO experiments for the sequential assignment and chemical shift analysis of intrinsically disordered proteins. J Biomol Nmr . 2012/05/15. 2012;53(2):139–48.

96. Novacek J, Janda L, Dopitova R, Zidek L, Sklenar V. Efficient protocol for backbone and side-chain assignments of large, intrinsically disordered proteins: transient secondary structure analysis of 49.2 kDa microtubule associated protein 2c. J Biomol Nmr. 2013/07/24. 2013;56(4):291–301. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23877929

97. Pons M. A "Russian Doll" Approach to More Efficient Acquisition of IDP NMR Spectra. Biophys J . 2019;1–2. Available from: https://doi.org/10.1016/j.bpj.2019.06.001

98. Plaxco KW, Morton CJ, Grimshaw SB, Jones JA, Pitkeathly M, Campbell LD, et al. The effects of guanidine hydrochloride on the "random coil" conformations and NMR chemical shifts of the peptide series GGXGG. J Biomol NMR. 1997;10(3):221–30.

99. Smith LJ, Bolin KA, Schwalbe H, MacArthur MW, Thornton JM, Dobson CM. Analysis of main chain torsion angles in proteins: Prediction of NMR coupling constants for native and random coil conformations. J Mol Biol. 1996;255(3):494–506.

100. Salmon L, Jensen MR, Bernadó P, Blackledge M. Measurement and Analysis of NMR Residual Dipolar Couplings for the Study of Intrinsically Disordered Proteins. In: Uversky VN, Dunker AK, editors. Intrinsically Disordered Protein Analysis: Volume 1, Methods and Experimental Tools . Totowa, NJ: Humana Press; 2012. p. 115–25. Available from: https://doi.org/10.1007/978-1-61779-927-3_9

101. Zweckstetter M, Bax A. Prediction of sterically induced alignment in a dilute liquid crystalline phase: Aid to protein structure determination by NMR. J Am Chem Soc. 2000;122(15):3791–2.

102. Crowhurst KA, Forman-Kay JD. Aromatic and methyl NOEs highlight hydrophobic clustering in the unfolded state of an SH3 domain. Biochemistry. 2003;42(29):8687–95.

103. Battiste JL, Wagner G. Utilization of site-directed spin labeling and high-resolution heteronuclear nuclear magnetic resonance for global fold determination of large proteins with limited nuclear overhauser effect data. Biochemistry. 2000;39(18):5355–65.

104. Dyson HJ, Wright PE. Unfolded Proteins and Protein Folding Studied by NMR. Chem Rev. 2004;104(8):3607–22.

105. Gillespie JR, Shortle D. Characterization of long-range structure in the denatured state of staphylococcal nuclease. I. Paramagnetic relaxation enhancement by nitroxide spin labels. J Mol Biol. 1997;268(1):158–69.

106. Bodenhausen G, Ruben DJ. Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. Chem Phys Lett. 1980;69(1):185–9.

107. Renner C, Schleicher M, Moroder L, Holak TA. Practical aspects of the 2D 15N-[1h]-NOE experiment. J Biomol Nmr. 2002/06/14. 2002;23(1):23–33. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12061715

108. Wright PE, Dyson HJ. Intrinsically Disordered Proteins in Cellular Signaling and Regulation. Nat Rev Mol Cell Biol. 2015;16(1):18–29.

109. Costello J, Castelli LM, Rowe W, Kershaw CJ, Talavera D, Mohammad-Qureshi SS, et al. Global mRNA selection mechanisms for translation initiation. Genome Biol. 2015 Jan 5;16(1):10. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25650959

110. Das S, Das B. eIF4G—an integrator of mRNA metabolism? FEMS Yeast Res. 2016 Oct 14;16(7). Available from: https://doi.org/10.1093/femsyr/fow087

111. Tarun Jr. SZ, Wells SE, Deardorff JA, Sachs AB. Translation initiation factor eIF4G mediates in vitro poly(A) tail-dependent translation. Proc Natl Acad Sci U S A. 1997/08/19. 1997;94(17):9046–51. Available from: https://www.ncbi.nlm.nih.gov/pubmed/9256432

112. Kroschwald S, Munder MC, Maharana S, Franzmann TM, Richter D, Ruer M, et al. Different Material States of Pub1 Condensates Define Distinct Modes of Stress Adaptation and Recovery. Cell Rep . 2018;23(11):3327–39. Available from: https://doi.org/10.1016/j.celrep.2018.05.041

113. Costello JL, Kershaw CJ, Castelli LM, Talavera D, Rowe W, Sims PFG, et al. Dynamic changes in eIF4F-mRNA interactions revealed by global analyses of environmental stress responses. Genome Biol . 2017 Oct 27;18(1):201. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29078784

CHAPTER 2

A CON-based NMR assignment strategy for pro-rich intrinsically disordered proteins with low signal dispersion: the C-terminal domain of Histone H1.0 as a case study

2.1. ABSTRACT

The C-terminal domain of histone H1.0 (C-H1.0) is involved in DNA binding and is a main determinant of the chromatin condensing properties of histone H1.0. Phosphorylation at the (S/T)-P-X-(K/R) motifs affects DNA binding and is crucial for regulation of C-H1.0 function. Since C-H1.0 is an intrinsically disordered domain, solution NMR is an excellent approach to characterize the effect of phosphorylation on the structural and dynamic properties of C-H1.0. However, it's very repetitive, low-amino acid-diverse and Pro-rich sequence, together with the low signal dispersion observed at the ^{1}H–^{15}N HSQC spectra of both non- and tri-phosphorylated C-H1.0 preclude the use of standard ^{1}H-detected assignment strategies. We have achieved an essentially complete assignment of the heavy backbone atoms (^{15}N, ^{13}C′ and ^{13}C$_\alpha$), as well as ^{1}H^N and ^{13}C$_\beta$ nuclei, of non- and tri-phosphorylated C-H1.0 by applying a novel ^{13}C-detected CON-based strategy. No C-H1.0 region with a clear secondary structure tendency was detected by chemical shift analyses, confirming at residue level that C-H1.0 is disordered in aqueous solution. Phosphorylation only affected the chemical shifts of phosphorylated Thr's, and their adjacent residues. Heteronuclear {^{1}H}–^{15}N NOEs were also essentially equal in the non- and tri-phosphorylated states. Hence, structural tendencies and dynamic properties of C-H1.0 free in aqueous solution are unmodified by phosphorylation. We propose that the assignment strategy used for C-H1.0, which is based on the acquisition of only a few 3D spectra, is an excellent choice for short-lived intrinsically disordered proteins with repetitive sequences.

Keywords: Intrinsically disordered domain · IDP · Phosphorylation · Histone · NMR assignment strategy

2.2. INTRODUCTION

H1 linker histones bind to linker DNA regions on the surface of the nucleosome, and participate in the maintenance of chromatin higher-order structure and in gene regulation. In mammals, histone H1 is a multigene family composed of seven somatic subtypes H1.1–H1.5, H1.0 and H1x [1]. Metazoan H1 contains three distinct domains: a short N-terminal domain (20–35 residues), a central globular domain (about 80 residues), and a long C-terminal domain (approximately 100 residues). Terminal domains are variable in length, amino acid sequence and post-translational modifications, and thus they determine subtype specificity [2,3]. The C-terminal domain is the primary determinant of H1 binding to chromatin in vitro and in vivo, and also is responsible for the preferential binding of H1 to scaffold-associated regions [4–6]. There is increasing evidence that H1 interacts with proteins of different functional categories, including pre-mRNA splicing, core histone chaperones and transcription-associated proteins [7]. It has also been described that the C-terminal domain interacts with specific proteins such as the apoptotic nuclease DFF40 [8] and prothymosin alpha [9]. In particular, the study of H1.0 interactome has revealed that one-third of the H1.0-dependent interaction proteins found in the nucleolus were mediated by the C-terminal domain [10].

H1 terminal domains are enriched in disorder-promoting amino acids, including lysine, serine, proline and alanine. Peptides derived from the basic portion of the N-terminal domain of subtypes H1.0 and H1.4, close to the globular domain, have been studied in aqueous solution and in presence of 2,2,2 trifluoroethanol (TFE) by circular dichroism (CD) and NMR [11–13]. The results revealed that while both peptides were disordered in aqueous solution, helical populations of 40–50% were detected in the presence of 90% TFE. A further study of the H1.0-derived peptides by Fourier Transform Infrared Spectroscopy (FTIR) showed that DNA induced an amount of α-helical structure equivalent to that observed in TFE [11,12]. Analysis of a peptide derived from the first portion of the C-terminal domain of H1.0, residues 99–121, provided similar results by NMR in presence of TFE and by FTIR in presence of DNA [12,14].

FTIR studies on the C-terminal domain of H1.0 (C-H1.0) showed that this domain was mostly disordered when free in aqueous solution, but acquired significant amounts of secondary structure when bound to DNA [15].

Thus, based on the structural information available up to now, H1 terminal domains can be classified as intrinsically disordered regions with coupled binding and folding. Further evidence of the folding of C-H1.0 upon binding to DNA and to reconstituted chromatin has been provided by FRET studies [16–18].

Cell-cycle dependent phosphorylation is the most studied post-translational modification of histone H1, as it may be involved in the regulation of chromatin dynamics [19]. H1 is phosphorylated by cyclin-dependent kinases (CDK) at the (S/T)-P-X-(K/R) consensus sequence, which are mainly located in the C-terminal domain. It has been described that phosphorylation of C-H1.0 by CDK2 triggers a conformational change, which results in significant changes in the secondary structure of the domain when bound to DNA [20]. Additional evidences of this conformational change in chromatin have been provided by FTIR and FRET [21,22]. Interestingly, some results suggest that the conformational change induced by phosphorylation is associated with *cis/trans* isomerization of the proline adjacent to the phosphorylated residues (Raghuram et al. 2013). NMR analyses of a 23-mer C-H1.0-derived peptide, which contains a CDK consensus sequence (TPKK), showed that the proline adjacent to the phosphoryl Thr was predominantly in trans conformation in the non-phosphorylated state [14]. In the presence of 90% of TFE, the TPKK sequence adopts a type I β-turn conformation, a σ-turn conformation or a combination of both, in fast equilibrium with unfolded states.

All the structural data about the whole C-H1.0 either free or DNA-bound has been obtained by techniques (CD, FTIR and FRET), which do not provide information at atomic level. In this context, we decided to use solution NMR to characterize the structural and dynamic properties of the full-length C-H1.0, and to examine the effect of phosphorylation. Considering that C-H1.0 is an intrinsically disordered domain, which has a very repetitive sequence composed of only a few types of amino acids, and contains a large number of Pro residues (see **Figure 1a, b**), the NMR assignment task was not expected to be easy. But, we did not anticipate to have so low signal dispersion as found at the 2D ^{1}H–^{15}N HSQC of C-H1.0 (**Figure 1c**). The low signal dispersion and high signal overlap observed in that spectrum, together with the repetitive sequence impeded the application of standard assignment strategies starting from the amide protons (see below). Therefore, we examined the alternative assignment strategies, which have been proposed in the last years.

The most successful ones are the ^{13}C-detected CON-based approaches, which correlate two consecutive CON groups in a protein [23–26]. Some of them are based on the acquisition of several 5D and/or 4D experiments recorded using non-uniform sampling [23,25,26]. These protocols speed up the assignment process providing backbone and aliphatic chemical shifts, but they require long acquisition times. As a result, they are not applicable when proteins are short-lived, because of degradation/stability problems. This was the case of C-H1.0 and pT-C-H1.0, which we found to get degraded in a few days. Hence, we searched for other assignment strategies, and succeeded by using a ^{13}C-detected CON-based approach, which uses only a few 3D spectra, which reduces the total acquisition time in relation with high-dimensionality approaches. This assignment strategy is described below, and proposed as an excellent alternative strategy in the case of IDP's with repetitive, low-amino acid diverse, and Pro-rich sequences and, in particular, if they are short-lived and signal dispersion at 2D ^{1}H–^{15}N HSQC is low, and signal overlap high.

The heavy atoms of the backbone (^{15}N, ^{13}C', and ^{13}C$_\alpha$), and the ^{1}H^N and ^{13}C$_\beta$ nuclei were fully assigned for C-H1.0 in aqueous solution in its non-phosphorylated state as well as tri-phosphorylated, denoted as pT-C-H1.0. Analyses of these chemical shifts indicated the absence of any secondary structure tendency in C-H1.0, either non-phosphorylated or tri-phosphorylated, when free in solution. Furthermore, dynamic characterization showed that both C-H1.0 and pT-C-H1.0 have a very high flexibility on the fast time scale.

2.3. RESULTS AND DISCUSSION

^{1}H-detected amide-based assignment strategies are not applicable to C-H1.0

For our solution NMR study, we used a C-H1.0 construct, whose overall secondary structure has been previously examined by FTIR [20]. This construct, denoted as C-H1.0, corresponds to residues 98–193 of mouse histone H1.0 (**Figure 1a**), and contains an additional N-terminal Met residue, two extra amino acids added by the restriction enzyme used for cloning, and a C-terminal 6xHis-tag, which are required for expression and purification of the protein. In terms of simplicity, herein we are going to number the C-H1.0 sequence from 1 (the N-terminal Met) to 105 (**Figure 1a**).

The sequence of C-H1.0 contains only 12 types of residues (**Figure 1b**). It lacks Gly residues and, excluding the six His from the histidine-tag, the only aromatic residue among them is a Phe (F10). Also, three out of the 12 types represent about a 65% of the sequence; they are Lys (38.1%; 40 out of 105), Ala (16.2%; 17 out of 105), and Pro (11.4%; 12 out of 105). As a consequence, many stretches of residues are repeated along the sequence, i.e. one 6-residue stretch ([21]ATPKKA[26] & [55]ATP-KKA[60]; **Figure 1a**), one 5-residue segment ([28]KPKKA[32] & [75]KPKKA[79]), and two 4-residue stretches ([39]KKPK[42] & [39]KKPK[42], and [40]KPKA[43] & [83]KPKA[86]) are each found twice in the C-H1.0 sequence; besides the 6- and 5-residue stretches share the 4-residue PKKA sub-sequence, which is four times in the sequence ([23]PKKA[26], [29]PKKA[32], [57]PKKA[60] and [76]PKKA[79]). Because of these repetitions, the residues of the same type, in particular the very numerous Lys, have very similar local chemical environments, which translates into their nuclei (particularly, $^{13}C_\alpha$ and $^{13}C_\beta$ carbons) having very similar chemical shifts, so that their cross-peaks appear at the same or very close places in the NMR spectra.

Furthermore, C-H1.0 is an intrinsically disorder protein, which shows very little dispersion in the $^{1}H^{N}$ nuclei; all amide protons appear in 0.51 ppm (**Figure 1c**). This range for amide protons is about a 25% narrower than that typically observed in prototype IDPs, such as α-synuclein (about a 1.0 ppm; 140 residues, BMRB-6968). In brief, the ^{1}H-detected conventional methods based on the triple resonance experiments, [27,28] in which assignment starts from the amide groups ($^{1}H^{N}$/^{15}N) are not suitable for C-H1.0 due to the severely overlapped signals.

As a first alternative, we intended to apply a ^{1}H-detected methodology using the experiment HNcacoNH, [29–32] which correlates the $^{1}H^{N}$ and ^{15}N chemical shifts of two consecutive amide moieties without involving other nuclei. However, the main drawback of these methods is that the connection path is broken at Pro residues. In our case, the sequence of C-H1.0 contains 12 Pro, so that the amide connection path splits into 13 segments, whose length is between 3 and 7 residues, except for the N- and C-terminal segments, which are 18 residues long. This sums up to the overlapping signals problem in the ^{1}H–^{15}N HSQC and the high redundant sequence. On the whole, using this approach we are only able to assign around a 67% of the amide HN groups of C-H1.0.
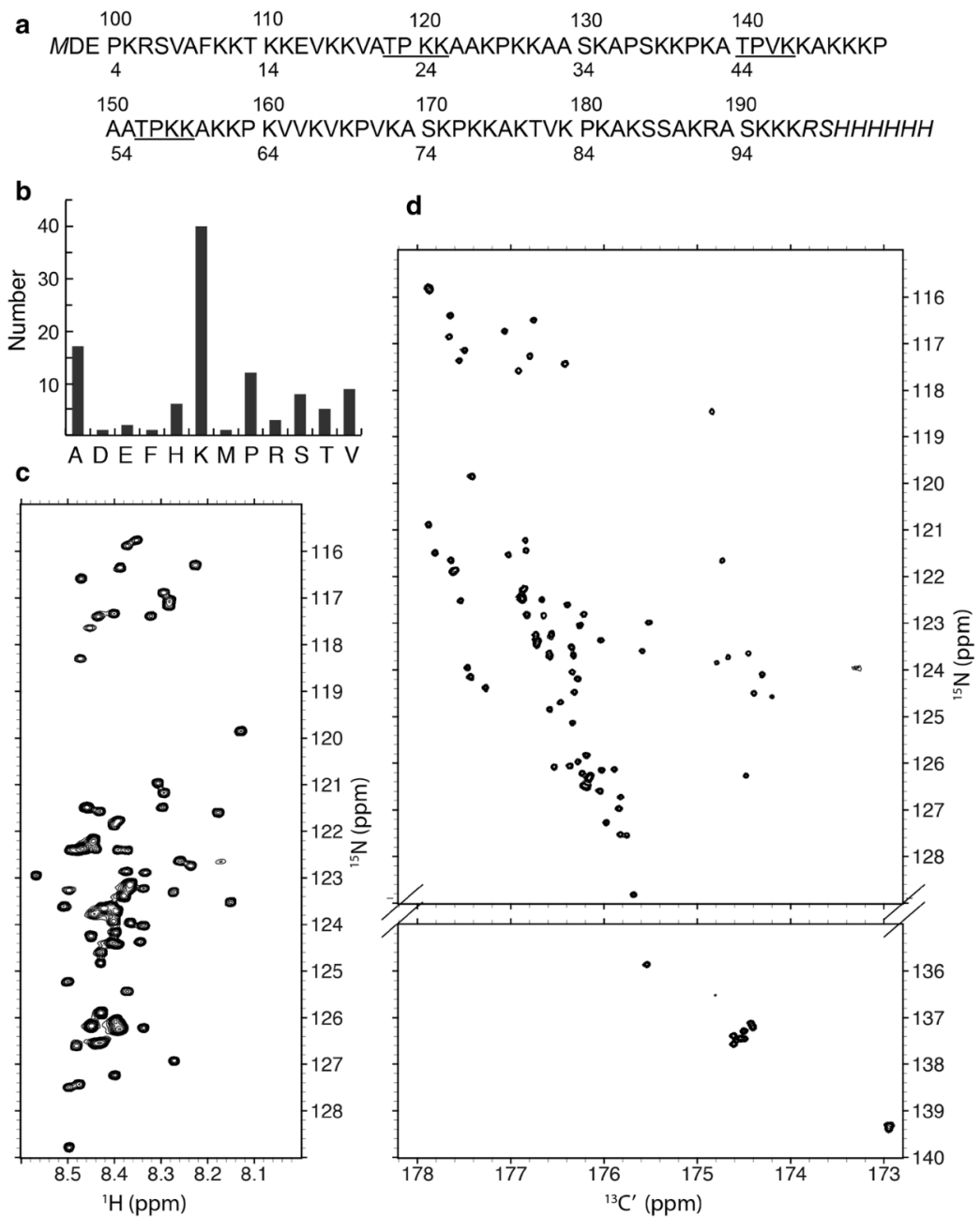
Figure 1. a) Sequence of C-H1.0 including the cloning-tag residues (in italics). Phosphorylation sequences are underlined. Numbers above the sequence correspond to full-length H1.0, and below to the construct used in this work. **b)** Bar plot showing the total number of each residue type in the C-H1.0 sequence. The absent residues are not included in the plot. **c)** 2D ^{1}H–^{15}N HSQC and **d)** 2D CON spectra recorded for C-H1.0 at 1 mM concentration in H$_2$O/ D$_2$O 9:1 v/v at pH 5.5 and 25 °C. The vertical y-axis of the upper part of the 2D CON is identical to that in the 2D ^{1}H–^{15}N HSQC

C-H1.0 assignment was achieved by a novel ^{13}C-detected CON-based strategy

Considering the failure to achieve a full or at least almost full assignment using the available ^{1}H-detected methodologies, we decided to explore alternative strategies based on ^{13}C-detected experiments. Although they are less sensitive than proton detection, we can benefit of its advantages relative to the wide chemical shift range of carbonyl carbons (^{13}C′), as well as the direct observation of cross-peaks for Pro residues. The higher dispersion of ^{13}C′–^{15}N cross-peaks relative to the ^{1}H^{N}-^{15}N cross-peaks can be appreciated by comparison of the 2D CON spectrum of C-H1.0 (**Figure 1d**) with the 2D ^{1}H–^{15}N HSQC spectrum acquired using the same C-H1.0 sample (**Figure 1c**).

In the 2D CON spectrum, the number of cross-peaks that we can count approximately coincides with those expected from the C-H1.0 sequence, and we observe signals for the Pro residues (see lower part of **Figure 1d**). The assignment strategies proposed up to now starting from 2D CON spectra consist in the use of ^{13}C-detected triple resonance experiments, [33–36] equivalent to those starting from the amide protons in the standard ^{1}H-detected strategy [27]. Thus, for sequential assignment they make use of nuclei with poorer dispersion (^{13}C$_\alpha$, ^{13}C$_\beta$ or ^{1}H$_\alpha$), which may adversely affect the assignment process, particularly in the case of repetitive sequences, such as C-H1.0. To overcome these difficulties some authors proposed the use of ^{1}H-detected [37,38] or ^{13}C-detected [23,26,39–42] high-dimensionality experiments, 4D or 5D. However, their application to C-H1.0 was problematic because of the sequence peculiarities of C-H1.0 (a sequence with low-amino acid diversity, many repeats, and a relatively high number of Pro residues; see above), together with its very low long-term stability. Therefore, we decided to explore the use of a simpler alternative.

Our proposal is based on the acquisition of a single 2D CON spectrum plus two 3D experiments: hacacoNcaNCO and hacaCOncaNCO, to directly correlate two consecutive CO–N groups in a protein, i.e., without mediation of the other nuclei [24]. Main advantage of these experiments is that the connection path is not interrupted at Pro residues, since the experiments use the ^{1}H$_\alpha$ protons as the starting point of the magnetization transfer pathway.

This is very important in Pro-rich proteins, such as C-H1.0, which contains 12 Pro (**Figure 1a, b**). To get additional information to determine the type of spin system, we collected a 3D CBCACON [33].

Thus, using only a set of four NMR experiments, i.e., 2D CON, 3D hacacoNcaNCO, 3D hacaCOncaNCO, and 3D CBCACON, we achieved an essentially complete assignment of the backbone ^{15}N, $^{13}C_\alpha$, and ^{13}C′ nuclei as well as the $^{13}C_\beta$ carbons of C-H1.0. The only unassigned residues correspond to K97, and the C-terminal cloning-tag (residues 98–105), as well as the $^{13}C_\alpha$ and $^{13}C_\beta$ carbons of K52 and K96. These ^{13}C and ^{15}N chemical shifts have been deposited in the BioMagRes-Bank database (http://www.bmrb.wisc.edu) under BMRB Accession Number 27537.

Tri-phosphorylated C-H1.0 was assigned by the ^{13}C-detected CON-based strategy

From the biological point of view, it is important to examine the structural and dynamic properties not only of C-H1.0, but also how they are affected by phosphorylation. Therefore, we proceeded to acquire NMR spectra of pT-C-H1.0, in which the Thr residues of the three phosphorylation motifs present in its sequence (^{22}TPKK^{25}, ^{44}TPVK^{47}, and ^{56}TPKK^{59}; **Fig. 1a**) are phosphorylated.

As in the case of non-phosphorylated C-H1.0, the 2D ^{1}H–^{15}N HSQC spectrum of tri-phosphorylated CH1.0 (denoted as pT-C-H1.0) showed low dispersion and high signal overlap, but most of the expected cross-peaks were present in the 2D CON of pT-C-H1.0 **(Fig. 2)**. It was not possible to unambiguously assign the cross-peaks in this spectrum by comparison to the previously assigned 2D CON spectrum of C-H1.0. Although the differences in chemical shift between C-H1.0 and pT-C-H1.0 resulted to be small **(see below)**, they suffice to impede the assignment of the phosphorylated protein just by comparison of CON spectra.

Therefore, we applied the CON- based strategy used for the non-phosphorylated protein **(see above)** to pT-C-H1.0, as a new test case for the validity of the method. As in the case of C-H1.0, an almost complete assignment of ^{15}N, $^{13}C_\alpha$, $^{13}C_\beta$ and ^{13}C′ nuclei was obtained by analyses of 3D hacacoNcaNCO, 3D hacaCOncaNCO, and 3D CBCACON acquired for pT-C-H1.0. In this case, the unassigned residues correspond to segment 100–105 of the C-terminal cloning-tag, the $^{13}C_\alpha$ and $^{13}C_\beta$ carbons of S89, K91, and S94, the ^{13}C′ carbon of R5, and the ^{15}N nitrogen of R6. These chemical shifts have been deposited in the BioMagResBank database (http://www.bmrb.wisc.edu) under BMRB Accession Number 27538.
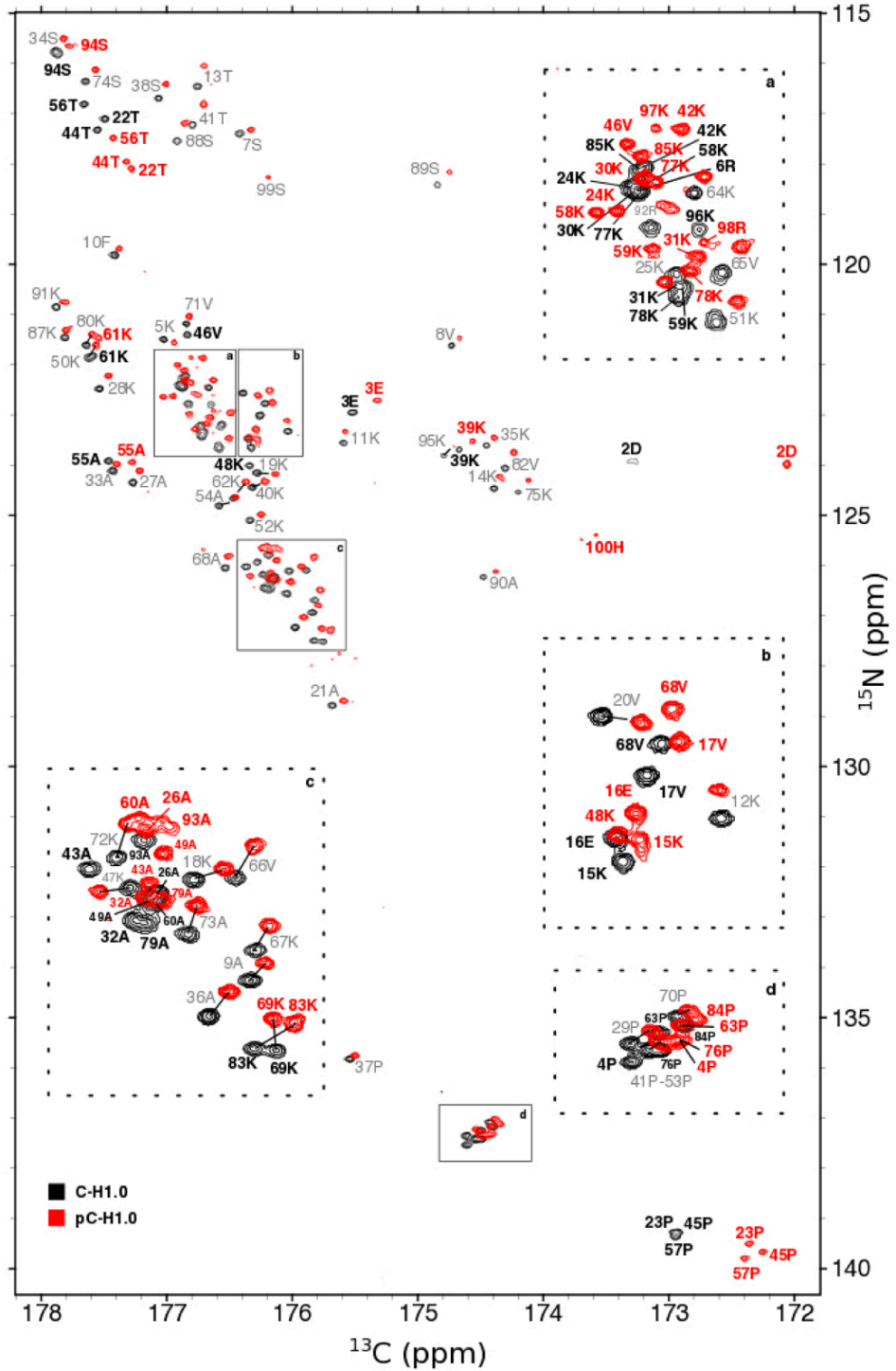
Figure 2. 2D CON spectra of 1 mM ^{15}N, ^{13}C C-H1.0 (black contours) and ^{15}N, ^{13}C pT-C-H1.0 (red contours) in H$_2$O/D$_2$O 9:1 v/v at pH 5.5 and 25 °C. Cross-peaks are labeled with the one letter code for amino acids and the residue number in the primary sequence according to the ^{15}N nucleus (see **Fig. 1a**; notice that each cross-peak corresponds to ^{15}N of residue i, and ^{13}C' of residue i − 1). Some crowd regions are zoomed for better visualization.

Chemical shift-based structural characterization of C-H1.0 and pT-C-H1.0

Once assigned C-H1.0 and pT-C-H1.0, we proceeded to see whether any region in these domains show a tendency towards some secondary structure (α-helical or β-sheet). It is well established that the regions without any secondary structure tendency show chemical shifts very close to reference values, whereas those with some tendency to be structured differ from reference values. The controversy in this point relates to the best reference values to be used. We are going to use those reported by Poulsen's group, [43,44] because they are generally considered as the most appropriate for intrinsically disordered proteins. Among the assigned chemical shifts, those of $^{13}C_\alpha$ carbons are probably the most reliable as indicators of secondary structure.

Therefore, we calculated the chemical shift deviations of $^{13}C_\alpha$ carbons ($\Delta\delta_{C\alpha} = \delta_{C\alpha}^{observed} - \delta_{C\alpha}^{reference}$, ppm) of C-H1.0, using the Poulsen's reference values, as implemented at Poulsen's server (https ://spin.niddk .nih.gov/bax/nmrse rver/Pouls en_rc_CS/), and plot them as a function of residue number (**Figure 3a**). The $\Delta\delta_{C\alpha}$ were all within the random coil range ($|\Delta\delta_{C\alpha}| \leq 0.4$ ppm), and no region could be pointed out as having even the slightest secondary structure tendency (**Figure 3a**). This conclusion was confirmed by examination of the chemical shift deviations for $^{13}C_\beta$ and $^{13}C'$ nuclei, in which no secondary structure tendency was detectable (**Suppl. Figure S1**). To check whether the use of other reference values could affect this result, we calculated the CSI (chemical shift index) as implemented at Wishart's server (http://csi3.wisha rtlab .com/ cgi-bin/index .php) [45]. The CSI uses a different set of random coil values, and takes into account $^{13}C_\beta$, $^1H^N$, and ^{15}N, apart from the $^{13}C_\alpha$ and $^{13}C'$, and gives a value of $+ 1$ for helical residues, $- 1$ for sheet residues, and zero for random coil residues. In C-H1.0, CSI was zero for all the residues, which indicates a fully unstructured polypeptide chain.

Next, we compare the chemical shifts of tri-phosphorylated pT-C-H1.0 relative to non-phosphorylated C-H1.0. **Figure 3b** shows the chemical shift perturbation (CSP), as defined in "Materials and methods", as function of residue number. The only significant differences are found for the three Thr residues, whose hydroxyl group is phosphorylated, and for their adjacent residues. Thus, it seems that phosphorylation has only local effects.

Circular dichroism (CD) data of C-H1.0 and pT-C-H1.0 also indicate that phosphorylation does not induce any conformational change (see **Suppl. Figure S2**). Nevertheless, we decided to check whether the chemical shift deviations for $^{13}C_\alpha$, $^{13}C_\beta$, and $^{13}C'$, as well as the CSI by themselves might show some secondary structure tendency, but the problem is that there is no appropriate chemical shift reference for phosphorylated Thr (pThr).

The only reference values available for pThr are those measured in a short peptide of sequence Ac-GGpTGG-NH$_2$, [46] and they can be inadequate for Pro-preceding pThr's. Using the Thr reference for pThr, the profiles of $\Delta\delta_{C\alpha}$, $\Delta\delta_{C\beta}$, and $\Delta\delta_{C'}$ **(Suppl. Fig. S1)** and a CSI value of zero for all the residues indicate that pT-C-H1.0 has no detectable secondary structure tendency, as in C-H1.0. The only significant differences placed around the pThr residues. As deduced from the CSP profile **(Fig. 3b)**, phosphorylation has only local effects.
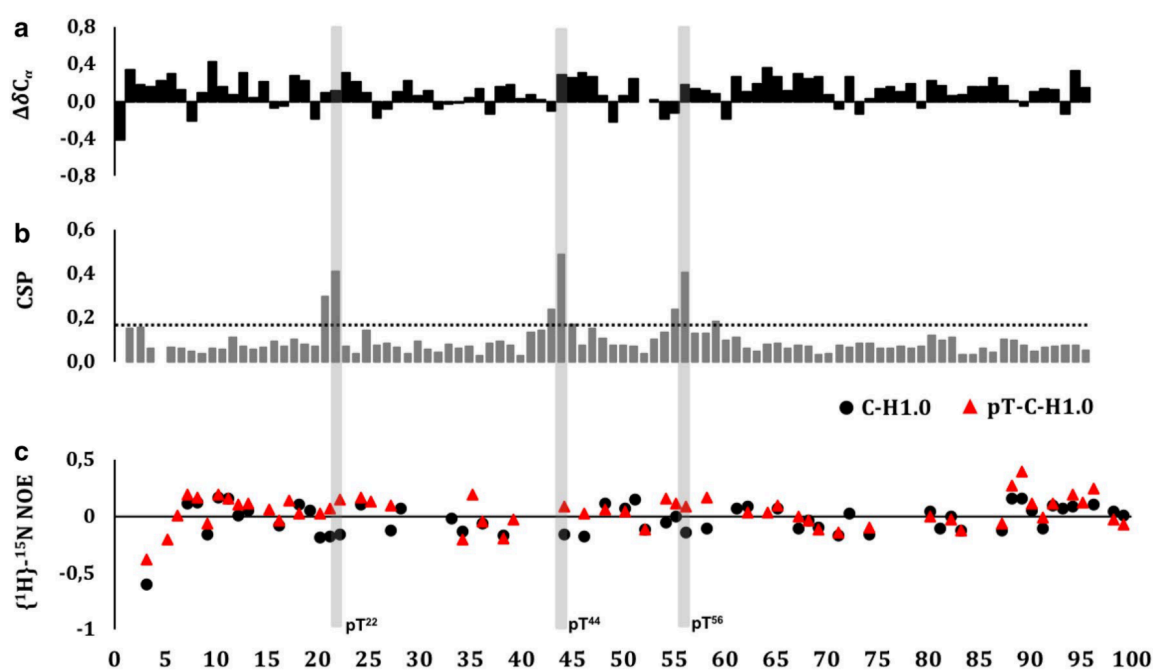
Figure 3. **a)** $\Delta\delta_{C\alpha}$ values ($\Delta\delta_{C\alpha} = \Delta\delta_{C\alpha}^{observed} - \Delta\delta_{C\alpha}^{reference}$, ppm) of C-H1.0 plotted as a function of residue number. Reference random coil values were taken from Poulsen (see "Methods"). **b)** Chemical shifts perturbation (CSP) due to phosphorylation (see "Methods"). The horizontal line is the averaged CSP plus the standard deviation. **c)** {^1H}–^{15}N heteronuclear NOEs for C-H1.0 (circles) and pT-C-H1.0 (triangles) in H$_2$O/D$_2$O 9:1 v/v at pH 5.5 and 25 °C. Grey vertical bars indicate the positions of the phosphorylated Thr residues.

We have also examined whether phosphorylation might affect the *cis/trans* isomerization equilibrium of the X-Pro bonds. To that end, we have analyzed the chemical shifts for the $^{13}C_\beta$ carbons of the 12 Pro residues. All of them are in the range characteristic of trans X-Pro bonds [47] in both the non-phosphorylated C-H1.0 and the tri-phosphorylated pT-C-H1.0. To get further confirmation of this result, we have run the PROMEGA program [48], which uses the chemical shifts of several nuclei (^{15}N, $^{13}C'$, $^{13}C_\alpha$ and $^{13}C_\beta$) from the Pro residue and its preceding and following residues, and takes into account amino acid sequence. The PROMEGA output indicates that all X-Pro bonds are trans in C-H1.0 and pT-C-H1.0, as deduced only on the basis of Pro $^{13}C_\beta$ chemical shifts. No minor signals attributable to low-populated *cis* species were detected.

Dynamics characterization of C-H1.0 and pT-C-H1.0

To obtain information of protein dynamics in the fast time scale (ps-ns), we collected heteronuclear $\{^1H\}$–^{15}N NOE experiments for C-H1.0 and pT-C-H1.0. To assign the signals at these spectra we need to extend the assignment to the $^1H^N$ protons. This was successfully done by analyses of the corresponding high-resolution 3D HNCO spectrum, which is very sensitive, and correlates the already assigned $^{13}C'$, and ^{15}N to the $^1H^N$ proton. To solve a few remaining ambiguities we used 3D HncacoNH and hNcacoNH experiments acquired using NUS. Out of the 92 $^1H^N$ protons present in C-H1.0 and pT-C-H1.0 (105 residues in total minus the N-terminal residue and minus the 12 Pro residues, which lack amide protons), we could assign 66 in C-H1.0, and 68 in pT-C-H1.0.

Once assigned these $^1H^N$ amide protons, we could measure the heteronuclear $\{^1H\}$-^{15}N NOE for 55 residues of C-H1.0 and for 56 residues in the tri-phosphorylated pT-C-H1.0. Those of the other residues, even if assigned, could not be measured reliably because of the high level of overlap. These NOEs, which have been deposited in the BioMagResBank database (http://www.bmrb.wisc.edu) under BMRB Accession Numbers 27537 and 27538, are plotted as a function of sequence in **Figure 3c**. The N-terminal residues, which show negative values, are slightly more flexible than the rest of the protein in both C-H1.0 and pT-C-H1.0. Excluding the N-terminal residues, the average values of these NOEs are 0.01 ± 0.11 for C-H1.0 and 0.07 ± 0.12 for pT-C-H1.0, which indicates that both proteins have a very high flexibility on the fast time scale (picoseconds to nanoseconds).

No significant differences can be appreciated between the non- and the tri-phosphorylated forms. Thus, flexibility in C-H1.0 is not affected by phosphorylation, whether this result can be extended to other IDPs or not is an open question.

2.4. CONCLUSIONS

The C-terminal domain of histone H1.0 (C-H1.0) plays a regulatory role in the functionality of histone H1.0, which is regulated by phosphorylation of Thr at motifs [118]TPKK[121], [140]TPVK[143] and [152]TPKK[155]. That this domain is intrinsically disordered hampers its structural characterization. NMR is probably the best technique to get structural information about IDPs, but it requires spectral assignment, which is more complex in IDPs than in globular proteins, because of signal dispersion decrease and signal overlap increase. In the case of C-H1.0, NMR assignment is still more complicated than in prototype IDPs because of its very repetitive sequence, its large amount of Pro residues, and a signal dispersion of only 0.5 ppm for the amide protons, which leads to strong signal overlap at the ^1H–^{15}N HSQC. This impedes the application of ^1H-detected assignment strategies starting at this experiment. Therefore, we applied a ^{13}C-detected assignment strategy based on the analyses of a set of only four NMR experiments, i.e., 2D CON, 3D hacacoNcaNCO, 3D hacaCOncaNCO, and 3D CBCACON. This strategy allow us to successfully assign the backbone ^{15}N, $^{13}C_\alpha$, and $^{13}C'$ nuclei as well as the $^{13}C_\beta$ carbons of C-H1.0 and its tri-phosphorylated form pT-C-H1.0. We propose this strategy as a best choice for NMR assignment in Pro-rich IDPs with repetitive sequences, which usually show very poor signal dispersion and a strong signal overlap in the ^1H–^{15}N HSQC spectra, particularly if they have short-lives.

The availability of the assignment has allowed us to perform the first structural and dynamics characterization of the non- and tri-phosphorylated C-H1.0, free in aqueous solution. Analyses of chemical shifts and heteronuclear {^1H}–^{15}N-NOEs indicated that the two forms have a similar structural and dynamics behavior; both lack of any region with detectable secondary structure tendency, and are very flexible. The effect of phosphorylation is local and affects only the chemical shifts of the phosphorylated Thr residues (T22, T44 and T56), and their neighbor residues (preceding and following). Regarding the *cis/trans* Pro isomerism, all X-Pro bonds are *trans* in both C-H1.0 and pT-C-H1.0 free in solution, so Thr phosphorylation does not affect the isomer state of the following

Pro. To have the assignment of C-H1.0 and pT-C-H1.0 opens the way to further characterization of these domains to understand the molecular bases of their biological relevance, such as DNA binding or interactions with other proteins.

Of particular interest would be the characterization of the secondary structure induced by DNA-binding, which in the phosphorylated forms might include *cis/trans* isomerization of prolines adjacent to phosphorylated residues. The application of the ^{13}C-detected CON-based assignment strategy described here, as well as the availability of the C-H1.0 chemical shifts would provide further details to the recently published NMR study of the interaction between ProTα and human histone H1.0, [49] in which ProTα was fully characterized but H1.0 was not assigned.

2.5. MATERIALS AND METHODS

Expression and purification of the C-terminal domain of H1.0 (C-H1.0)

The coding sequence of the C-terminal domain of mouse H1.0 was cloned in the pQE-60 vector (Qiagen), with a 6xHis-tag at the C-terminus, as previously described [6]. The recombinant expression vector pCTH1.0 was transformed into *E. coli* M15 (Qiagen). The isotopically labeled C-H1.0 was prepared by Marley et al.'s method [50]. Briefly, cells were grown to an OD$_{600nm}$ of 0.8 in rich medium (Luria–Bertani), centrifuged at 5000 g for 15 min and re-suspended in minimal medium M9 salts. After a second step of centrifugation the cell mass obtained in 2L of rich medium was finally re-suspended in 1L of minimal medium (M9) prepared with ^{13}C-D-glucose (Cambridge Isotope Laboratories Inc.) and ^{15}NH$_4$Cl (Cambridge Isotope Laboratories Inc.). Following a short period for growth recovery, protein expression was induced with 1 mM IPTG, allowing expression to proceed for 4 h at 37 °C. Cells were then harvested and, if necessary, stored at − 80 °C. Next, cells were lysed in the lysis buffer (0.05 M NaH$_2$PO$_4$, 0.75 M NaCl, 0.02 M imidazol) plus 4 M guanidine hydrochloride, pH 8.0 for 15 min at room temperature. The extract was centrifuged at 20,000 g for 25 min. The supernatants were loaded on a HiTrap chelating HP column (GE Healthcare) equilibrated with lysis buffer. The column was then washed in three steps with lysis buffer containing increasing amounts of imidazol: 40, 60 and 80 mM. Finally, the proteins were eluted with 250 mM imidazol in lysis buffer and desalted by gel filtration through Sephadex G-25 (Sigma-Aldrich).

Preparation of the phosphorylated C-terminal domain of H1.0 (pT-C-H1.0)

The isotopically labeled C-H1.0 was phosphorylated in vitro using CDK2-cyclin A2 kinase (Sigma-Aldrich) as previously described [20]. Briefly, the phosphorylation reaction was carried out in 50 mM Tris–HCl, 10 mM $MgCl_2$, 1 mM EGTA, 20 mM dithiotreitol, pH 7.5, plus 200 µM ATP and 1U of CDK2-cyclin A per 5 µg of C-H1.0. The mixture was incubated at 30 °C for 1 h and the reaction buffer was eliminated by gel filtration on a HiTrap desalting column (GE Healthcare). The extent of phosphorylation was evaluated by MALDI-TOF mass spectrometry.

NMR experiments

All NMR experiments were collected at 298 K on a Bruker Avance spectrometer, operating at a ^{1}H frequency of 800.1 MHz equipped with a TCI cryoprobe. [13]C and ^{15}N uniformly labeled C-H1.0 and pT-C-H1.0 samples were prepared at approximately 1 mM protein concentration in H_2O/D_2O 9:1 v/v at pH 5.5. The set of collected experiments for each protein consisted of four ^{13}C-detected experiments, i.e. 2D CON, and 3D hacacoNcaNCO, hacaCOncaNCO and CBCANCO, and two [1]H-detected experiments, i.e., 2D ^{1}H–^{15}N HSQC and 3D HNCO. In the case of pT-C-H1.0 3D HncacoNH and hNcacoNH were also recorded. In the case of 3D spectra, the less sensitive ^{13}C-detected experiments were recorded using linear sampling, and the most sensitive [1]H-detected using non-uniform sampling (NUS).

The 2D ^{13}C-detected CON experiment was acquired with 512 and 256 complex points in the [13]C′ direct dimension and ^{15}N indirect dimension. The carriers of [13]C′ and ^{15}N dimensions were set at 172.5 ppm and 127.5 ppm, respectively, and the spectral widths were 10 and 27 ppm in the [13]C′ and ^{15}N dimensions, respectively. The total measurement time was 1.8 h. The 3D ^{13}C-detected spectra were acquired with 512 complex points in the [13]C′ direct dimension (f3), 24 complex points in the indirect dimension [15]N (f2) attached directly to [13]C′, 24 complex points in the indirect dimension [13]C′ or [15]N (f1) and 48 complex point in [13]$C_β$/[13]$C_α$ (f1) dimension for CBCANCO. The carriers of the [13]C′ and ^{15}N dimension were set at 172.5 ppm and 127.5 ppm, respectively. The spectral widths were 10 and 27 ppm in the [13]C′ and ^{15}N dimensions, respectively, or 39 ppm in the ^{13}C dimension of CBCANCO. The total measurement time was 30 h (16 scans) or 19 h (8 scans) for CBCANCO experiment.

For the 2D ^{1}H–^{15}N HSQC spectrum 1K and 256 complex data points were acquired for direct and indirect dimension, respectively, four scans were accumulated and the total experimental time was 41 min. The 3D ^{1}H^N-detected HNCO experiments were acquired with 1K complex points in the direct dimension (f3) and 24 complex points in the ^{15}N dimension (f2) and 64 complex points in the ^{13}C dimension (f1), and 8 scans. The total measurement time was 8.5 h. The 3D HncacoNH and hNcacoNH experiments were acquired with non-uniform sampling (NUS). They were acquired using a 310-complex point sampling schedule with 8 scans per FID in 3 h and 40 min each one. The maximum increment in the NUS schedule is 48 or 32 for the ^{15}N attached directly to ^{1}H acquired in the direct dimension (f2) and ^{15}N or ^{1}H (f1) dimensions, respectively. ^{15}N (f2) is centered in 122 ppm with the spectral width being 1297.33 Hz, and ^{15}N and ^{1}H (f1) are centered in 122 ppm and 7.17 ppm with the spectral width being 1297.33 Hz and 3880.5 Hz, respectively. The directly observed ^{1}H dimension has a spectral width of 8802.817 Hz (centered at 4.75 ppm) with 1024 complex points.

{^{1}H}–^{15}N NOE experiments for C-H1.0 and pT-C-H1.0 were acquired in an interleaved manner with and without proton saturation during an overall recycling delay of 10 s to ensure the maximal development of NOEs before acquisition and to allow solvent relaxation, thus avoiding transfer of saturation to the most exposed amide protons of the protein between scans [51]. They were acquired with 1 K and 128 complex data points for direct and indirect dimension, respectively, 8 scans were accumulated and the total experimental time was 14 h.

For all the experiments, the resulting matrix was zero filled to double the number of original points in all dimensions and shifted squared sine-bell apodization functions were applied in all dimensions prior to Fourier transformation. Spectra were processed using either TOPSPIN v2.1 pl6 (Bruker, Inc) or NMRpipe [52] and the istHMS reconstruction method was used to process NUS data [53]. Finally, they were analyzed with the programs SPARKY (T. Goddard and D.G. Kneller, SPARKY 3, University of California, San Francisco, USA) and/or NMRview [54]. Chemical shift perturbation (CSP) were obtained by applying the following equation: CSP = [(δ_N $^{pT-C-H1.0}$ – δ_N $^{C-H1.0}$)^2 + (($\delta_{C'}$ $^{pT-C-H1.0}$ – $\delta_{C'}$ $^{C-H1.0}$) A)^2]$^{1/2}$ where the scaling factor A, which is the ratio between ^{15}N and ^{13}C' spectral widths (A = $SW_N/SW_{C'}$), is equal to 0.37 [55].

2.6. ACKNOWLEDGMENTS

2.7. REFERENCES

1.	Talbert PB, Ahmad K, Almouzni G, Ausio J, Berger F, Bhalla PL, et al. A unified phylogeny-based nomenclature for histone variants. Epigenetics Chromatin. 2012/06/02. 2012;5:7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22650316

2.	Roque A, Ponte I, Suau P. Post-translational modifications of the intrinsically disordered terminal domains of histone H1: effects on secondary structure and chromatin dynamics. Chromosoma. 2016/04/22. 2017;126(1):83–91. Available from: http://www.ncbi.nlm.nih.gov/ pubmed/27098855

3.	Roque A, Ponte I, Suau P. Interplay between histone H1 structure and function. Biochim Biophys Acta. 2015/09/30. 2016;1859(3):444–54. Available from: http:// www.ncbi.nlm.nih.gov/pubmed/26415976

4.	Th'ng JP, Sung R, Ye M, Hendzel MJ. H1 family histones in the nucleus. Control of binding and localization by the C-terminal domain. J Biol Chem. 2005/05/25. 2005;280(30):27809–14. Available from: http:// www.ncbi.nlm.nih.gov/pubmed/15911621

5.	Hendzel MJ, Lever MA, Crawford E, Th'ng JP. The C-terminal domain is the primary determinant of histone H1 binding to chromatin in vivo. J Biol Chem. 2004/02/27. 2004;279(19):20028–34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14985337

6.	Roque A, Orrego M, Ponte I, Suau P. The preferential binding of histone H1 to DNA scaffold-associated regions is determined by its C-terminal domain. Nucleic Acids Res. 2004/11/25. 2004;32(20):6111–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/ 15562002

7.	Kalashnikova AA, Rogge RA, Hansen JC. Linker histone H1 and protein-protein interactions. Biochim Biophys Acta . 2015/10/13. 2016;1859(3):455–61. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26455956

8.	Widlak P, Kalinowska M, Parseghian MH, Lu X, Hansen JC, Garrard WT. The histone H1 C-terminal domain binds to the apoptotic nuclease, DNA fragmentation factor (DFF40/CAD) and stimulates DNA cleavage. Biochemistry. 2005/05/25. 2005;44(21):7871–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed /15910001

9.	Zakharova NI, Sokolov V V, Suvorova AA, Shiau AI, Wu CL, Efstaf'eva AG. [Prothymosin alpha interacts with C-terminal domain of histone H1 and dissociates p53-histone H1 complex]. Mol Biol. 2011/10/01. 2011;45(4):679–88. Available from: http://www.ncbi.nlm.nih.gov/ pubmed/21954601

10. Kalashnikova AA, Winkler DD, McBryant SJ, Henderson RK, Herman JA, DeLuca JG, et al. Linker histone H1.0 interacts with an extensive network of proteins found in the nucleolus. Nucleic Acids Res . 2013/02/26. 2013;41(7):4026–35. Available from: http:// www.ncbi.nlm.nih.gov/pubmed/23435226

11. Vila R, Ponte I, Collado M, Arrondo JL, Suau P. Induction of secondary structure in a COOH-terminal peptide of histone H1 by interaction with the DNA: an infrared spectroscopy study. J Biol Chem. 2001/06/20. 2001;276(33):30898–903. Available from: http:// www.ncbi.nlm.nih.gov/pubmed/11413144

12. Vila R, Ponte I, Collado M, Arrondo JL, Jimenez MA, Rico M, et al. DNA-induced alpha-helical structure in the NH2-terminal domain of histone H1. J Biol Chem. 2001/10/05. 2001;276(49):46429–35. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11584004

13. Vila R, Ponte I, Jimenez MA, Rico M, Suau P. An inducible helix-Gly-Gly-helix motif in the N-terminal domain of histone H1e: a CD and NMR study. Protein Sci. 2002/01/16. 2002;11(2):214–20. Available from: http:// www.ncbi.nlm.nih.gov/pubmed/11790831

14. Vila R, Ponte I, Jimenez MA, Rico M, Suau P. A helix-turn motif in the C-terminal domain of histone H1. Protein Sci. 2000/05/04. 2000;9(4):627–36. Available from: http://www.ncbi.nlm.nih.gov/pubmed /10794405

15. Roque A, Iloro I, Ponte I, Arrondo JL, Suau P. DNA-induced secondary structure of the carboxyl-terminal domain of histone H1. J Biol Chem. 2005/07/12. 2005;280(37):32141–7. Available from: http:// www.ncbi.nlm.nih.gov/pubmed/16006555

16. Caterino TL, Fang H, Hayes JJ. Nucleosome linker DNA contacts and induces specific folding of the intrinsically disordered H1 carboxyl-terminal domain. Mol Cell Biol . 2011/04/06. 2011;31(11):2341–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21464206

17. Fang H, Clark DJ, Hayes JJ. DNA and nucleosomes direct distinct folding of a linker histone H1 C-terminal domain. Nucleic Acids Res. 2011/10/25. 2012;40(4):1475–84. Available from: http://www.ncbi.nlm.nih.gov /pubmed/22021384

18. Fang H, Wei S, Lee TH, Hayes JJ. Chromatin structure-dependent conformations of the H1 CTD. Nucleic Acids Res . 2016/11/02. 2016;44(19):9131–41. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27365050

19. Liao R, Mizzen CA. Interphase H1 phosphorylation: Regulation and functions in chromatin. Biochim Biophys Acta . 2015/12/15. 2016;1859(3):476–85. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26657617

20. Roque A, Ponte I, Arrondo JL, Suau P. Phosphorylation of the carboxy-terminal domain of histone H1: effects on secondary structure and DNA condensation. Nucleic Acids Res. 2008/07/18. 2008;36(14):4719–26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/ 18632762

21. Lopez R, Sarg B, Lindner H, Bartolome S, Ponte I, Suau P, et al. Linker histone partial phosphorylation: effects on secondary structure and chromatin condensation. Nucleic Acids Res. 2015/04/15. 2015;43(9):4463–76. Available from: http://www.ncbi.nlm.nih.gov/ pubmed/25870416

22. Raghuram N, Strickfaden H, McDonald D, Williams K, Fang H, Mizzen C, et al. Pin1 promotes histone H1 dephosphorylation and stabilizes its binding to chromatin. J Cell Biol. 2013/10/09. 2013;203(1):57–71.

23. Novacek J, Zawadzka-Kazimierczuk A, Papouskova V, Zidek L, Sanderova H, Krasny L, et al. 5D 13C-detected experiments for backbone assignment of unstructured proteins with a very low signal dispersion. J Biomol Nmr . 2011/03/23. 2011;50(1):1–11. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21424579

24. Pantoja-Uceda D, Santoro J. New 13C-detected experiments for the assignment of intrinsically disordered proteins. J Biomol Nmr. 2014/04/05. 2014;59(1):43–50. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24699834

25. Bermel W, Felli IC, Gonnelli L, Kozminski W, Piai A, Pierattelli R, et al. High-dimensionality 13C direct-detected NMR experiments for the automatic assignment of intrinsically disordered proteins. J Biomol Nmr . 2013/11/10. 2013;57(4):353–61. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24203099

26. Dziekanski P, Grudziaz K, Jarvoll P, Kozminski W, Zawadzka-Kazimierczuk A. (13)C-detected NMR experiments for automatic resonance assignment of IDPs and multiple-fixing SMFT processing. J Biomol Nmr . 2015/04/24. 2015;62(2):179–90. Available from: http:// www.ncbi.nlm.nih.gov/pubmed/25902761

27. Sattler M, Schleucher J, Griesinger C. Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. Prog Nucl Magn Reson Spectrosc. 1999;34(2):93–158.

28. Permi P, Annila A. Coherence transfer in proteins. Prog Nucl Magn Reson Spectrosc. 2004;44(1–2):97–137.

29. Grzesiek S, Anglister J, Ren H, Bax A. C-13 Line Narrowing by H-2 Decoupling in H-2/C-13/N-15-Enriched Proteins - Application to Triple-Resonance 4d J-Connectivity of Sequential Amides. J Am Chem Soc. 1993;115(10):4369–70.

30. Panchal SC, Bhavesh NS, Hosur R V. Improved 3D triple resonance experiments, HNN and HN(C)N, for H-N and N-15 sequential correlations in (C-13, N-15) labeled proteins: Application to unfolded proteins. J Biomol Nmr. 2001;20(2):135–47.

31. Pantoja-Uceda D, Santoro J. Aliasing in reduced dimensionality NMR spectra: (3,2)D HNHA and (4,2)D HN(COCA)NH experiments as examples. J Biomol Nmr . 2009/10/24. 2009;45(4):351–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19851713

32. Sun ZY, Frueh DP, Selenko P, Hoch JC, Wagner G. Fast assignment of 15N-HSQC peaks using high-resolution 3D HNcocaNH experiments with non-uniform sampling. J Biomol Nmr. 2005/10/14. 2005;33(1):43–50. Available from: http://www.ncbi.nlm.nih.gov/pubmed/ 16222556

33. Bermel W, Bertini I, Duma L, Felli IC, Emsley L, Pierattelli R, et al. Complete assignment of heteronuclear protein resonances by protonless NMR spectroscopy. Angew Chem Int Ed Engl. 2005/04/16. 2005;44 (20):3089–92. Available from: http://www.ncbi.nlm.nih.gov/ pubmed/15832397

34. Bermel W, Bertini I, Csizmok V, Felli IC, Pierattelli R, Tompa P. H-start for exclusively heteronuclear NMR spectroscopy: the case of intrinsically disordered proteins. J Magn Reson . 2009/03/25. 2009;198(2):275–81. Available from: http://www.ncbi.nlm.nih.gov/ pubmed/19307141

35. Mantylahti S, Aitio O, Hellman M, Permi P. HA-detected experiments for the backbone assignment of intrinsically disordered proteins. J Biomol Nmr. 2010/05/04. 2010;47(3):171–81.

36.   Mantylahti S, Hellman M, Permi P. Extension of the HA-detection based approach: (HCA)CON(CA)H and (HCA)NCO(CA)H experiments for the main-chain assignment of intrinsically disordered proteins. J Biomol Nmr. 2011/01/25. 2011;49(2):99–109. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21259120

37.   Zerko S, Byrski P, Wlodarczyk-Pruszynski P, Gorka M, Ledolter K, Masliah E, et al. Five and four dimensional experiments for robust backbone resonance assignment of large intrinsically disordered proteins: application to Tau3x protein. 2016/07/20. Vol. 65, Journal of Biomolecular Nmr. 2016. p. 193–203. Available from: http://www.ncbi.nlm.nih.gov/ pubmed/27430223

38.   Piai A, Hosek T, Gonnelli L, Zawadzka-Kazimierczuk A, Kozminski W, Brutscher B, et al. "CON-CON" assignment strategy for highly flexible intrinsically disordered proteins. J Biomol Nmr. 2014/10/20. 2014 ;60(4):209–18. Available from: http://www.ncbi.nlm.nih.gov/ pubmed/25326659

39.   Novacek J, Haba NY, Chill JH, Zidek L, Sklenar V. 4D non-uniformly sampled HCBCACON and (1)J(NCalpha)-selective HCBCANCO experiments for the sequential assignment and chemical shift analysis of intrinsically disordered proteins. J Biomol Nmr. 2012/05/15. 2012;53(2):139–48. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22580891

40.   Novacek J, Janda L, Dopitova R, Zidek L, Sklenar V. Efficient protocol for backbone and side-chain assignments of large, intrinsically disordered proteins: transient secondary structure analysis of 49.2 kDa microtubule associated protein 2c. J Biomol Nmr. 2013/07/24. 2013;56(4):291–301. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23877929

41.   Bermel W, Bertini I, Felli IC, Gonnelli L, Kozminski W, Piai A, et al. Speeding up sequence specific assignment of IDPs. J Biomol Nmr. 2012/06/12. 2012;53(4):293–301. Available from: http://www.ncbi.nlm.nih.gov/ pubmed/22684679

42.   Motackova V, Novacek J, Zawadzka-Kazimierczuk A, Kazimierczuk K, Zidek L, Sanderova H, et al. Strategy for complete NMR assignment of disordered proteins with highly repetitive sequences based on resolution-enhanced 5D experiments. J Biomol Nmr. 2010/10/05. 2010;48(3):169–77. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20890634

43.   Kjaergaard M, Poulsen FM. Sequence correction of random coil chemical shifts: correlation between neighbor correction factors and changes in the Ramachandran distribution. J Biomol Nmr. 2011/05/24. 2011;50(2):157–65. Available from: http://www.ncbi.nlm.nih.gov/ pubmed/21604143

44.   Kjaergaard M, Brander S, Poulsen FM. Random coil chemical shift for intrinsically disordered proteins: effects of temperature and pH. J Biomol Nmr . 2011/01/15. 2011;49(2):139–49. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21234644

45.   Hafsa NE, Arndt D, Wishart DS. CSI 3.0: a web server for identifying secondary and super-secondary structure in proteins using NMR chemical shifts. Nucleic Acids Res. 2015/05/17. 2015;43(W1):W370-7. Available from: http://www.ncbi.nlm.nih.gov/ pubmed/25979265

46.   Bienkiewicz EA, Lumb KJ. Random-coil chemical shifts of phosphorylated amino acids. J Biomol Nmr . 2000/03/11. 1999;15(3):203–6. Available from: http://www.ncbi.nlm.nih.gov/ pubmed/10677823

47.   Schubert M, Labudde D, Oschkinat H, Schmieder P. A software tool for the prediction of Xaa-Pro peptide bond conformations in proteins based on 13C chemical shift statistics. J Biomol Nmr. 2002/12/24. 2002;24(2):149–54. Available from: http://www.ncbi.nlm.nih.gov/ pubmed/12495031

48.   Shen Y, Bax A. Prediction of Xaa-Pro peptide bond conformation from sequence and chemical shifts. J Biomol Nmr. 2009/12/31. 2010;46(3):199–204. Available from: http://www.ncbi.nlm.nih.gov/pubmed /20041279

49.   Borgia A, Borgia MB, Bugge K, Kissling VM, Heidarsson PO, Fernandes CB, et al. Extreme disorder in an ultrahigh-affinity protein complex. Nature. 2018/02/22. 2018;555(7694):61–6. Available from: http:// www.ncbi.nlm.nih.gov/pubmed/29466338

50.   Marley J, Lu M, Bracken C. A method for efficient isotopic labeling of recombinant proteins. J Biomol Nmr . 2001/06/30. 2001;20(1):71–5. Available from: http:// www.ncbi.nlm.nih.gov/pubmed/11430757

51.   Renner C, Schleicher M, Moroder L, Holak TA. Practical aspects of the 2D 15N-[1h]-NOE experiment. J Biomol Nmr. 2002/06/14. 2002;23(1):23–33. Available from: http://www.ncbi.nlm.nih.gov/pubmed /12061715

52.   Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol Nmr. 1995/11/01. 1995;6(3):277–93. Available from: http:// www.ncbi.nlm.nih.gov/pubmed/8520220

53.   Hyberts SG, Frueh DP, Arthanari H, Wagner G. FM reconstruction of non-uniformly sampled protein NMR data at higher dimensions and optimization by distillation. J Biomol Nmr . 2009/08/26. 2009;45(3):283–94. Available from: http://www.ncbi.nlm.nih.gov/ pubmed/ 19705283

54.   Johnson BA, Blevins RA. NMR View: A computer program for the visualization and analysis of NMR data. J Biomol Nmr. 1994/09/01. 1994;4(5):603–14. Available from: http://www.ncbi.nlm.nih.gov/pubmed/ 22911360

55.   Williamson MP. Using chemical shift perturbation to characterise ligand binding. Prog Nucl Magn Reson Spectrosc . 2013/08/22. 2013;73:1–16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23962882

2.8. SUPPORTING INFORMATION

Expression of the C-terminal domain of H1.0 (C-H1.0)

Plasmids corresponding to wild-type C-terminal histone H1.0 protein were transformed in *E. coli* M15 cells [1] as described in Material and Methods section. To see if we could enhance expression levels, we transformed the plasmid in *E. coli* BL21 (DE3) competent cells expressing in kanamycin containing (30 ug/l) LB medium. As a result, M15 and BL21 (DE3) *E. Coli* cells expressed the same level of protein, so we could use both.

On the other hand, the isotopically labeled media, described in Material and Methods, were prepared by Marley et al.'s protocol [2], that is, using M9 minimal medium with ^{13}C-D-glucose, ^{15}NH$_4$Cl and allowing expression to proceed for 4 h at 37 ºC. However, in trying to improve the expression level we tested K-MOPS derived minimal medium [3] supplemented with ^{15}NH$_4$Cl (1 g/l) and ^{13}C-D-glucose (4 g/l). In this case, the culture was grown at 25 ºC overnight. Approximately, the same expression levels were obtained, so C-H1.0 could be prepared using any of the two assayed minimal media types.

Purification of the C-terminal domain of H1.0 (C-H1.0)

The first protein sample, which had been purified following the procedure described in Materials and Methods, was degraded in less than one day. In case this degradation were caused in some way by the purification protocol, we decided to check the use of different purification steps, and see if we could establish a better purification method. The alternative protocol is the following: First, the cell pellet were resuspended in the lysis buffer (0.05 M NaH$_2$PO$_4$, 0.75 M NaCl, 0.02 M imidazol) plus 4 M guanidine hydrochloride pH 8.0, during 30 min at 4 ºC, instead of 15 min at room temperature. Then, it was sonicated (3 min ON, 10 min OFF and 30 % Amplitude) and the extract was centrifuged at 15000 rpm for 15 min. The supernatant was purified by metal affinity chromatography using HiTrap HP column. Then, the column was washed using 0.04 M imidazole only once. Part of the protein is lost if three washes are done, as in the protocol described in Material and Methods. Finally, the protein was eluted using the elution buffer (0.05 M NaH2PO4, 0.75 M NaCl, 0.25 M imidazol).

All purification process was done at 4 ºC. In this alternative protocol, a step using Sephadex G25 column was not done, because the purified protein was very clean, without degradation or aggregation after the previous step (see gel at Figure S1).
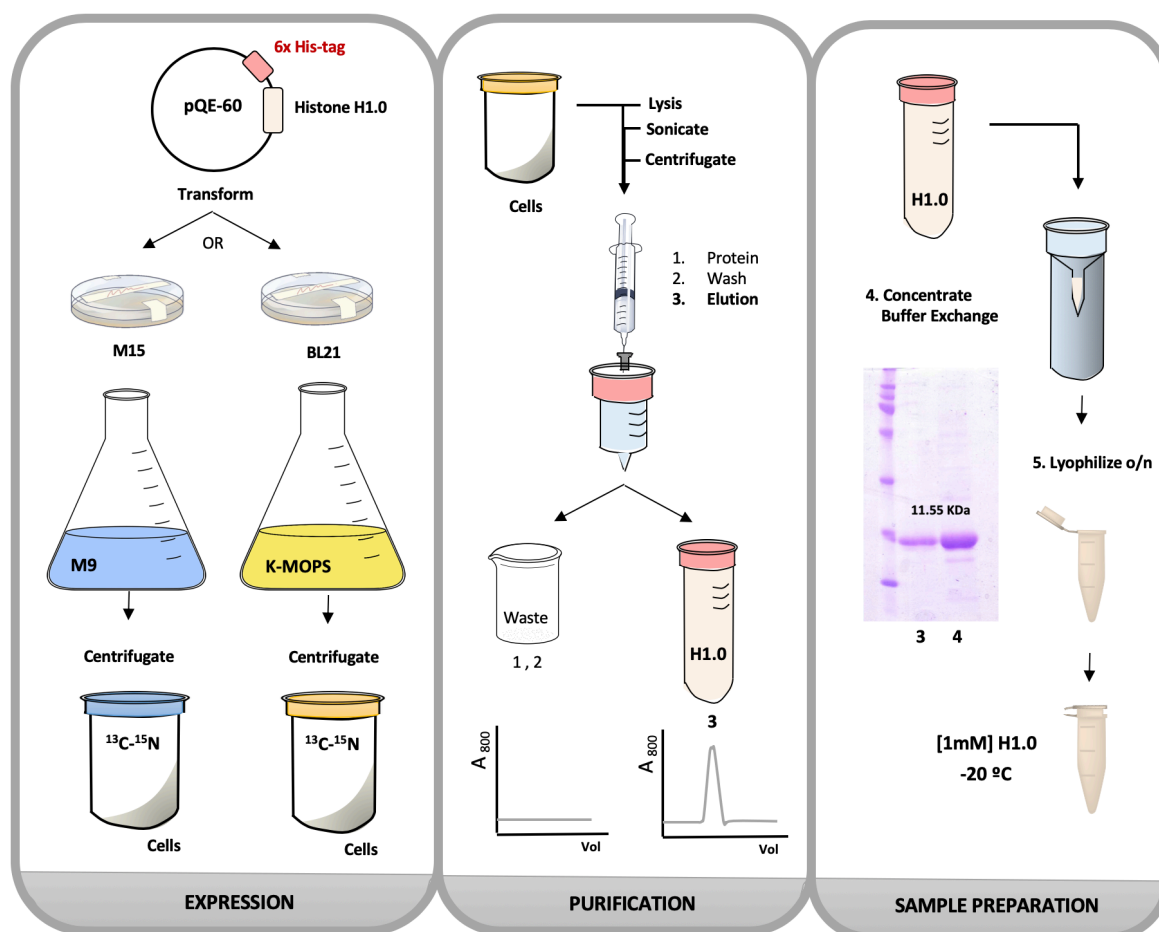
Figure S1. Graphical view of the two procedures of expression, purification and sample preparation followed for the non-phosphorylated C-terminal domain of histone H1.0 (C-H1.0).

Sample preparation of the C-terminal domain of H1.0 (C-H1.0)

The protein was concentrated using a Vivaspin 20 5 KDa MWCO concentrator (Sigma Aldrich) and the buffer exchanged to H_2O using the same concentrator or Nap^{TM}-5 Column Sephadex^{TM} G-25 DNA Grade (GE Healthcare). At the end, the sample was lyophilized to weigh it and calculate its concentration to prepare de NMR sample. The concentration of C-H1.0 could not be measured by UV

absorbance at 280 nm, which is the standard method to measure protein concentrations, because the sequence of C-H1.0 does not contain aromatic residues (see Figure 1a-b). ^{13}C and ^{15}N uniformly labeled C-terminal H1.0 sample was prepared at approximately 1 mM protein concentration in H_2O/D_2O 9:1 v/v at pH 5.5.

Phosphorylation of the C-terminal domain of H1.0 (C-H1.0)

Once the protein is purified, lyophilized and estimated its concentration, the protein would be ready to be phosphorylated. The complete phosphorylation of C-terminal domain of H1.0 in the three specific threonine's (T_{22}, T_{44} and T_{56}) is done using the enzyme CDK2-cyclin A as described in Materials and Methods section. The following Figure S2 represents the graphical steps of phosphorylation.
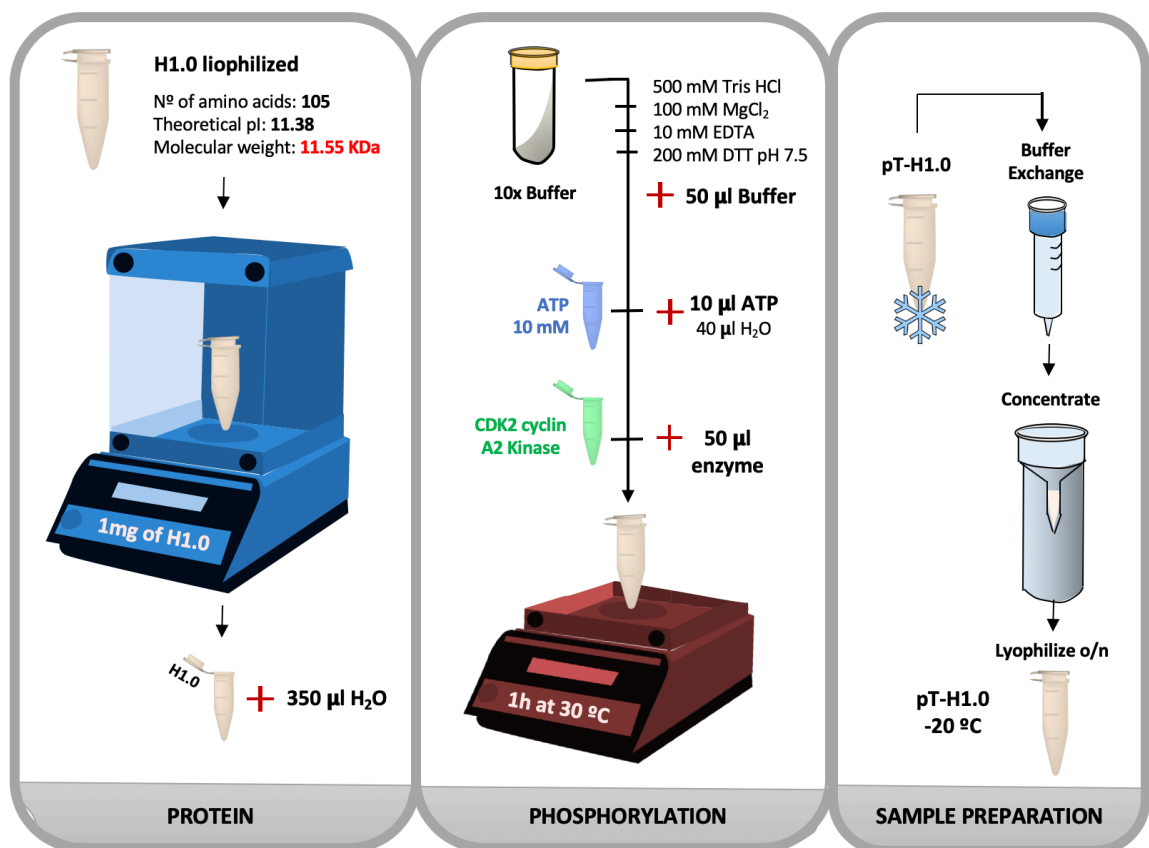

Figure S2. Graphical view of the procedure followed to phosphorylate the C-terminal domain of histone H1.0.

NMR experiments

The set of collected experiments recorded for C-terminal phosphorylated and non-phosphorylated Histone H1.0 was ^{13}C-detected experiments. These experiments (2D CON, 3D hacacoNcaNCO, hacaCOncaNCO and CBCANCO) made it possible to assign the protein. Two [1]H-detected experiments (2D ^{1}H-^{15}N HSQC and 3D HNCO) were also recorded.

Figure S3 shows the assignment strategy proposed in this paper. It starts from a 2D CON spectra [4–7], where the observed [13]C'-^{15}N correlations are between the [13]C' of residue i and the ^{15}N of the following residue i+1 ([13]C' i - [15]N i+1). As an example in the C- H1.0 sequence, we took the **57[13]C'-58[15]N** correlation (represented in red, Figure S3a), and labeled in the 2D CON shown in Figure S3b. To identify the next residue, that is, the **58[13]C'-59[15]N** correlation, and make possible the sequential assignment, the 3D hacaCOncaNCO spectrum is needed [8]. In this spectrum at the plane corresponding to the [13]C' i - [15]N i+1 correlation, we can find a cross-peak at the [13]C' i+1 chemical shift (ppm). In our example, we would obtain the **58[13]C'** (see top spectrum at Figure S3b).

To complete the next pair of residues, a 3D hacacoNcaNCO is required [8]. In this spectrum at the same plane used before ([13]C' i - [15]N i+1 correlation) we can find a cross-peak at the [15]N i+1 chemical shif (ppm). Therefore, this spectrum uses the 2D CON information to give us the complete new pair **58[13]C'-59[15]N**. However, these three spectra are not enough to assign the protein, because the sequence is very repetitive and has a large number of the same type of residues (Figure 1a-b). Therefore, we used a 3D experiment CBCACON [4], which for every CON signal ([13]C' i - [15]N i+1) gives the chemical shifts of the [13]C_α and [13]C_β of the residue i. Based on these chemical shifts, it is possible to determine the type of spin system, and hence solve ambiguities. In the example, at the CBCACON plane corresponding to the **57[13]C'-58[15]N** CON cross-peak, we find the **[13]C_α** and **[13]C_β** chemical shifts for P57 (see right spectrum at Figure S3b).

[1]H-detected strategy [9] could not be used in the case of C-H1.0. However, once assigned the 2D CON spectra, the [1]H_N amide chemical shifts could straightforwardly assigned in a 3D HNCO (Figure S4).
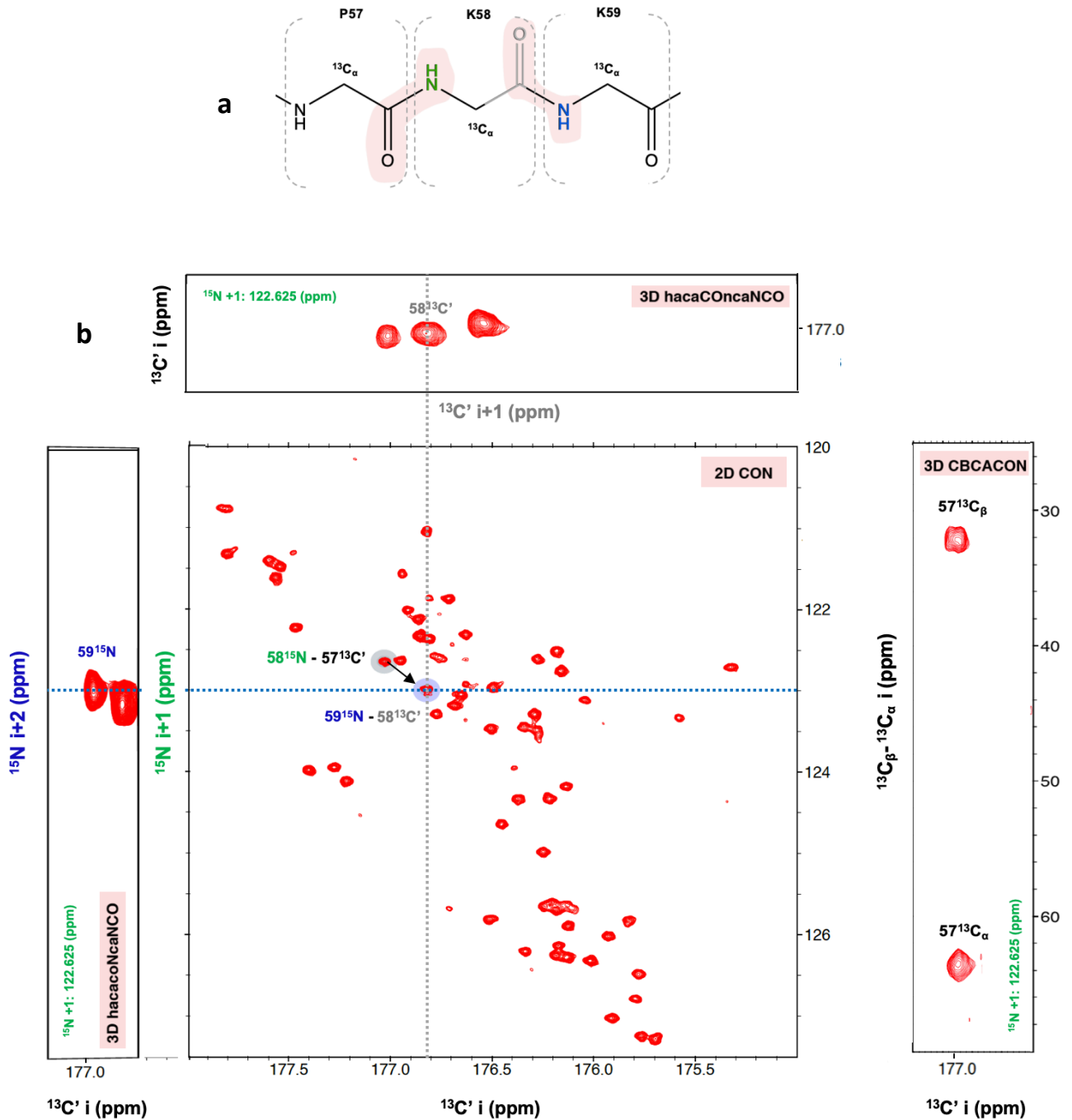
Figure S3. ^{13}C-detected CON-based assignment strategy. (a) Backbone of residues 57-59 of C-H1.0 (P57, K58 and K59 are residues i, i+1 and i+2, respectively). The atoms involved in the two 2D CON ^{13}C'-^{15}N correlations (57^{13}C'-58^{15}N and 59^{13}C'-60^{15}N) are represented in red. (b) **^{13}C' i - ^{15}N i+1** correlation of P57 in ppm (2D CON spectrum), **^{13}C' i - ^{15}N i+1** and **^{13}C$_\alpha$ i ^{13}C$_\beta$ i** carbons information to determinate the spin system of P57 (3D CBCACON), **^{13}C' i - ^{15}N i+1** with the new **^{13}C' i+1** position of K58 in ppm (3D hanaCOncaNCO spectrum), and **^{13}C' i - ^{15}N i+1** with the new **^{15}N i+2** position of K59 residue in ppm (3D hacacoNcaNCO).

Then, the assignment of the $^{1}H_N$ and ^{15}N amide chemical shifts can be transferred to the cross-peaks in the 2D $^{1}H,^{15}N$ HSQC experiment. Only some cross-peaks in very crowdy regions could not be assigned.
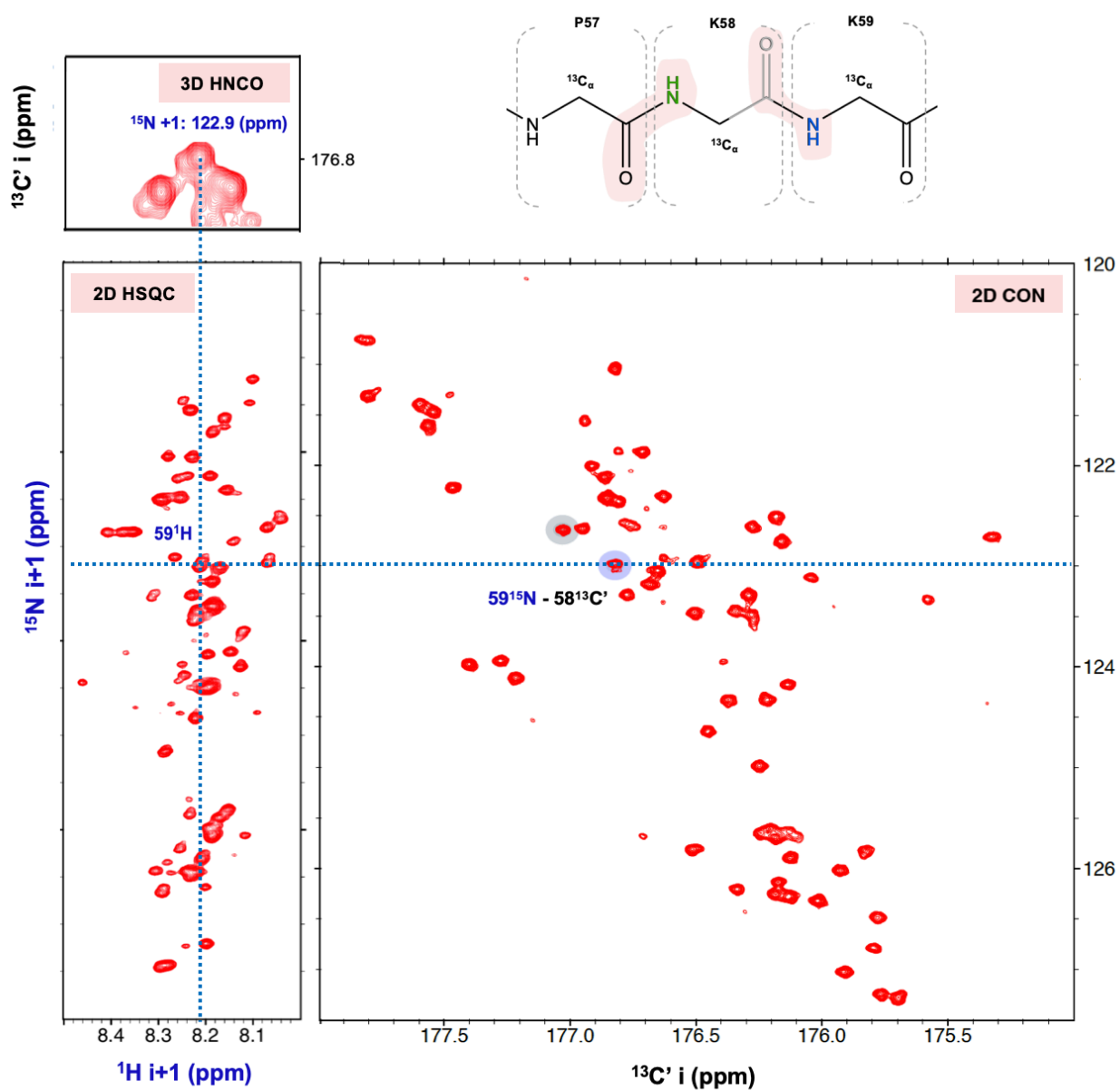
Figure S4. Assignment of 2D $^{1}H,^{15}N$-HSQC spectra in the ^{13}C-detected CON-based strategy. **$^{13}C'$ i** - **^{15}N i+1** correlation of K58 in ppm (2D CON spectrum), **$^{13}C'$ i** - **^{15}N i+1** with the new **^{1}H i+1** information of K59 residue (3D HNCO), and **^{15}N i+1** - **^{1}H i+1** correlation of K59 residue in ppm (2D HSQC).
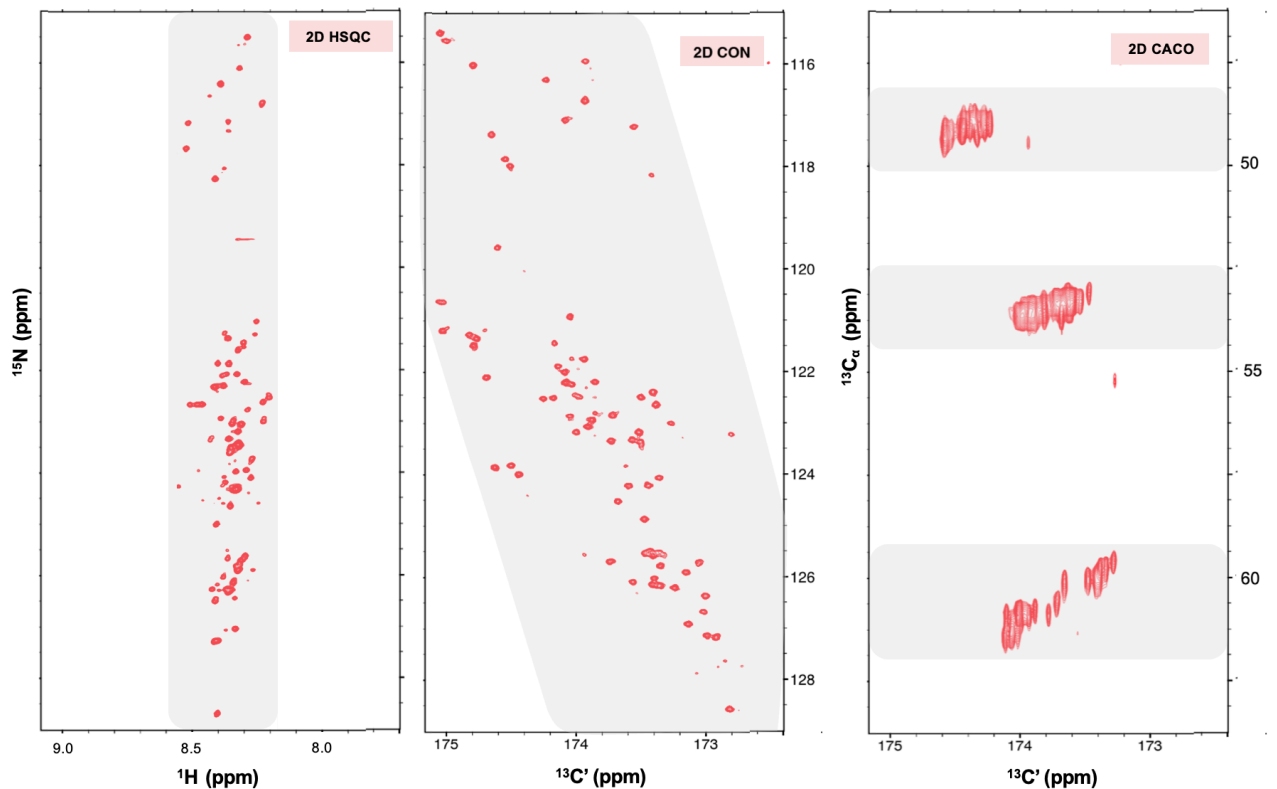
Figure S5. D ^{1}H-^{15}N HSQC, 2D ^{13}C'-^{15}N CON and 2D ^{13}C'-^{13}C$_\alpha$CACO spectra recorded for C terminal Histone H1.0 at 1 mM concentration in H$_2$O/D$_2$O 9:1 v/v at pH 5.5 and 25 ºC. Both, 2D ^{1}H-^{15}N HSQC and ^{13}C'-^{13}C$_\alpha$CACO signals show a lot of overlap. The 2D ^{13}C'-^{15}N CON signals have greater dispersion, which makes their assignment possible.
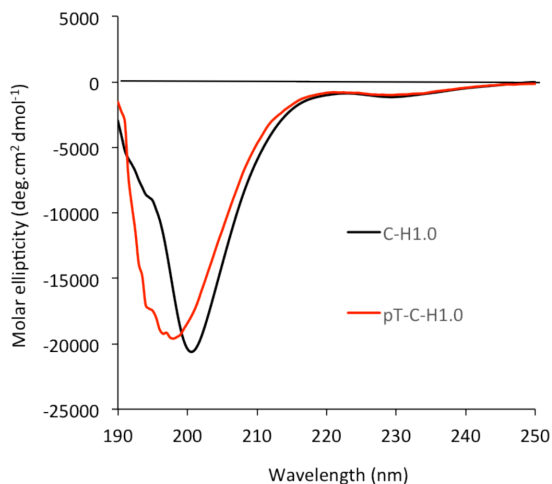
Figure S6. Far-UV CD spectra of non-phosphorylated (C-H1.0) and tri-phosphorylated (pT-C-H1.0) C-terminal domain of Histone H1.0. Protein samples were in 10 mM phosphate buffer at pH 7.0 plus 10 mM NaCl. CD spectra were recorded using a Jasco J-715 spectropolarimeter in 1 mm cuvettes at 20 ºC, and expressed as molar ellipticity.
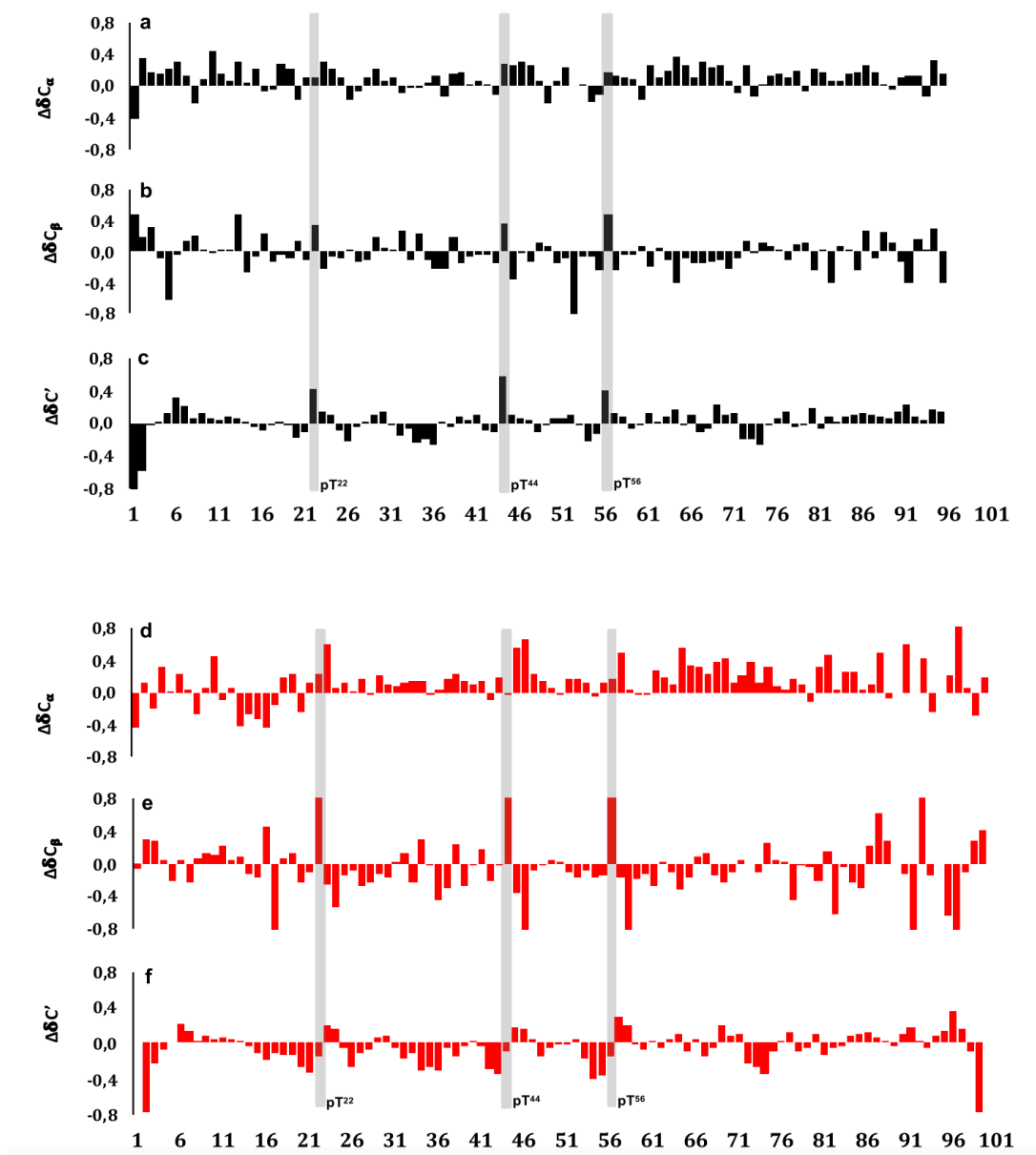
Figure S7. **Chemical shift deviations as a function of residue number for non- phosphorylated (C-H1.0; black bars) and tri-phosphorylated (pT-C-H1.0; red bars) C-terminal domain of histone H1.0:** $\Delta\delta_{C\alpha}$ values ($\Delta\delta_{C\alpha} = \Delta\delta_{C\alpha}^{observed} - \Delta\delta_{C\alpha}^{reference}$, ppm) of C-H1.0 **(panel a)** and pT-C-H1-0 **(panel d)**; $\Delta\delta_{C\beta}$ values ($\Delta\delta_{C\beta} = \Delta\delta_{C\beta}^{observed} - \Delta\delta_{C\beta}^{reference}$, ppm) of C-H1.0 **(panel b)** and pT-C-H1-0 **(panel e)**; and $\Delta\delta_{C'}$ values ($\Delta\delta_{C'} = \Delta\delta_{C'}^{observed} - \Delta\delta_{C'}^{reference}$, ppm) of C-H1.0 **(panel c)** and pT-C-H1-0 **(panel f)**. Note that $\Delta\delta$ for pT residues were calculated using the T reference values, so that the large $\Delta\delta_{C\beta}$ values observed for pT **(panel e)** are very likely due to inadequacy of the reference value.

2.9. SUPPORTING REFERENCES

1.   Roque A, Orrego M, Ponte I, Suau P. The preferential binding of histone H1 to DNA scaffold-associated regions is determined by its C-terminal domain. Nucleic Acids Res. 2004/11/25. 2004;32(20):6111–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15562002

2.   Marley J, Lu M, Bracken C. A method for efficient isotopic labeling of recombinant proteins. J Biomol Nmr. 2001/06/30. 2001;20(1):71–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11430757

3.   Neidhardt FC, Bloch PL, Smith DF. Culture medium for enterobacteria. J Bacteriol. 1974;119(3):736–47.

4.   Bermel W, Bertini I, Duma L, Felli IC, Emsley L, Pierattelli R, et al. Complete assignment of heteronuclear protein resonances by protonless NMR spectroscopy. Angew Chem Int Ed Engl. 2005/04/16. 2005; 44(20):3089–92. Available from: http://www.ncbi.nlm.nih.gov/pubmed/ 15832397

5.   Bermel W, Bertini I, Csizmok V, Felli IC, Pierattelli R, Tompa P. H-start for exclusively heteronuclear NMR spectroscopy: the case of intrinsically disordered proteins. J Magn Reson . 2009/03/25. 2009;198(2):275–81. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19307141

6.   Mantylahti S, Aitio O, Hellman M, Permi P. HA-detected experiments for the backbone assignment of intrinsically disordered proteins. J Biomol Nmr. 2010/05/04. 2010;47(3):171–81. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20437194

7.   Mantylahti S, Hellman M, Permi P. Extension of the HA-detection based approach: (HCA)CON(CA)H and (HCA)NCO(CA)H experiments for the main-chain assignment of intrinsically disordered proteins. J Biomol Nmr. 2011/01/25. 2011;49(2):99–109. Available from: http://www.ncbi.nlm.nih.gov/ pubmed/21259120

8.   Pantoja-Uceda D, Santoro J. New 13C-detected experiments for the assignment of intrinsically disordered proteins. J Biomol Nmr. 2014/04/05. 2014;59(1):43–50. Available from: http://www.ncbi.nlm.nih.gov/ pubmed/ 24699834

9.   Sattler M, Schleucher J, Griesinger C. Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. Prog Nucl Magn Reson Spectrosc. 1999;34(2):93–158.

CHAPTER 3

NMR studies on DNA recognition of the intrinsically disordered C-terminal domain of Histone H1.0

3.1   ABSTRACT

Linker histones play an important role in establishing and maintaining chromatin higher-order structure and in gene regulation. H1 Histones can act in transcriptional regulation through the modulation of chromatin condensation. The C-terminal domain of Histone H1.0 (C-H1.0) has been reported to bind to linker DNA, neutralizing the negative charge of phosphates and facilitating this chromatin condensation. C-H1.0 is phosphorylated by CDK2 in a cell cycle-dependent manner in the S/T-P-X-R/K motifs. Previous NMR studies have demonstrated that C-H1.0 is an intrinsically disordered domain and that phosphorylation has not effect on secondary structure or in dynamic properties. However, spectroscopic studies suggested that C-H1.0 experiences some structuration upon DNA interaction. Therefore, we performed a NMR study to get structural details into the C-H1.0 / DNA interaction at residue level. DNA titration of C-H1.0 was followed using 2D ^{13}C-detected CON spectra, which shows a good signal dispersion in contrast to the large signal overlap present in the ^{1}H,^{15}N-HSQC. Upon DNA titration the cross-peaks at the 2D CON spectra suffer changes in chemical shift and in intensity, which are not equal along C-H1.0 sequence. Both changes indicate that DNA interaction occurs preferentially at the C-terminal region. According to S^2 order parameters predicted from chemical shifts, this C-terminal region might become slightly less flexible. Regarding the phosphorylation motifs, their DNA affinity follows the order $T^{152} > T^{138} > T^{118}$, that is, depends on their proximity to the C-terminal region. On the other hand, chemical shift deviations indicated that DNA binding induces no secondary structure in any C-H1.0 region. Concerning the effect of phosphorylation, DNA interaction with the tri-phosphorylated pT-C-H1.0 is weaker than with the non-phosphorylated C-H1.0, but likely it also occurs throughout the C-terminal region. Interestingly, the behavior of the middle phosphorylation motif, $pT^{138}PVK$, upon DNA binding differs from those of the $pT^{118}PKK$ and $pT^{152}PKK$ motifs, and also from that of T^{138} in C-H1.0.

Keywords: NMR, Histone H1.0, DNA titration, Phosphorylation

79

3.2   INTRODUCTION

In the eukaryotic organisms DNA is found in the chromatin at the cell nucleus [1,2]. There, DNA binds to basic proteins, known as Histones to form the nucleosome [3]. H1 linker Histones, which binds to linker DNA regions on the surface of the nucleosome, are thought to be responsible for chromatin condensation and hence to have a regulatory role in transcription [4]. Three domains are distinguished in the sequence of linker histones: a short amino-terminal domain (NTD; 20–35 amino acids), a central globular domain (GD; ~80 amino acids), and a long carboxy-terminal domain (CTD; ~100 amino acids) (**Figure 1A**) [5]. The N- and C-terminal domains behave as intrinsically disordered, which, according to FT-IR studies, increase their regular secondary structure (α-helix and β-structure) content upon DNA-binding [6–9].

Mammalian H1 Histone has multiple subtypes [10,11], which differ in chromatin affinity, genomic localization, expression pattern and post-translational modifications (PTMs) [12]. The CTD domain seems to be responsible for the distinct chromatin-binding affinities of each H1 subtype [13–16]. Chromatin binding may also be modulated by post-transcriptional modifications (PTMs), in particular by those modifying the positive net charge of the CTD, like acetylation and phosphorylation [9].

Thus, the CTD of Histone H1 contains conserved short linear motifs (SLiMs), the consensus (S/T)-P-X-(K/R) sequences, which are recognized by cyclin-dependent kinases (CDKs) [17,18], and phosphorylated at the S/T residue. The levels of phosphorylation vary along the different phases of the cell cycle [19], such as Interphase and Mitosis. Along the Interphase the cell is prepared to division in three steps, such as G1, S and G2 phase. The level of Histone H1 phosphorylation is the lowest in G1 phase and rises during S and G2 phases [20,21]. The phosphorylation in the C-terminal domain of the H1.0 subtype (C-H1.0) occurs at three residues Thr118, Thr140 and Thr152 which will be referred to herein as "T^{22} T^{44} and T^{56}" (**Figure 1B**) [22]. Phosphorylation of C-H1.0 by CDK2 has a moderate effect on the affinity of this C-H1.0 for the DNA, and also affects its DNA-aggregating capacity. C-H1.0 phosphorylation is associated with chromatin relaxation and a decrease of H1 residence time [23,24]. Based on FTIR analysis, phosphorylation decreases the α-helix content and increases the β-structure in DNA-bound C-H1.0 [25,26]. Phosphorylation-triggered conformational change was also observed in full-length H1 Histone by FTIR in chromatin and by FRET analysis of reconstituted nucleosomes [23,27].

However, our recent NMR studies have evidenced that free C-H1.0 is a highly disordered domain, both non-phosphorylated and phosphorylated [28]. Nevertheless, details of a possible conformational change in C-H1.0 caused by DNA interaction at atomic level are unknown. Therefore, we decided to continue our NMR characterization of C-H1.0 by analysing its interaction with DNA and the effect that phosphorylation has on this interaction. Considering that at Interphase (G1, S or G2) in the cell (Figure 1C), non-phosphorylated and phosphorylate H1.0 species coexist and compete for DNA binding, we examined the effect of phosphorylation on DNA interaction using a sample containing equal amounts of non-phosphorylated C-H1.0 and triphosphorylated (pT-C-H1.0).
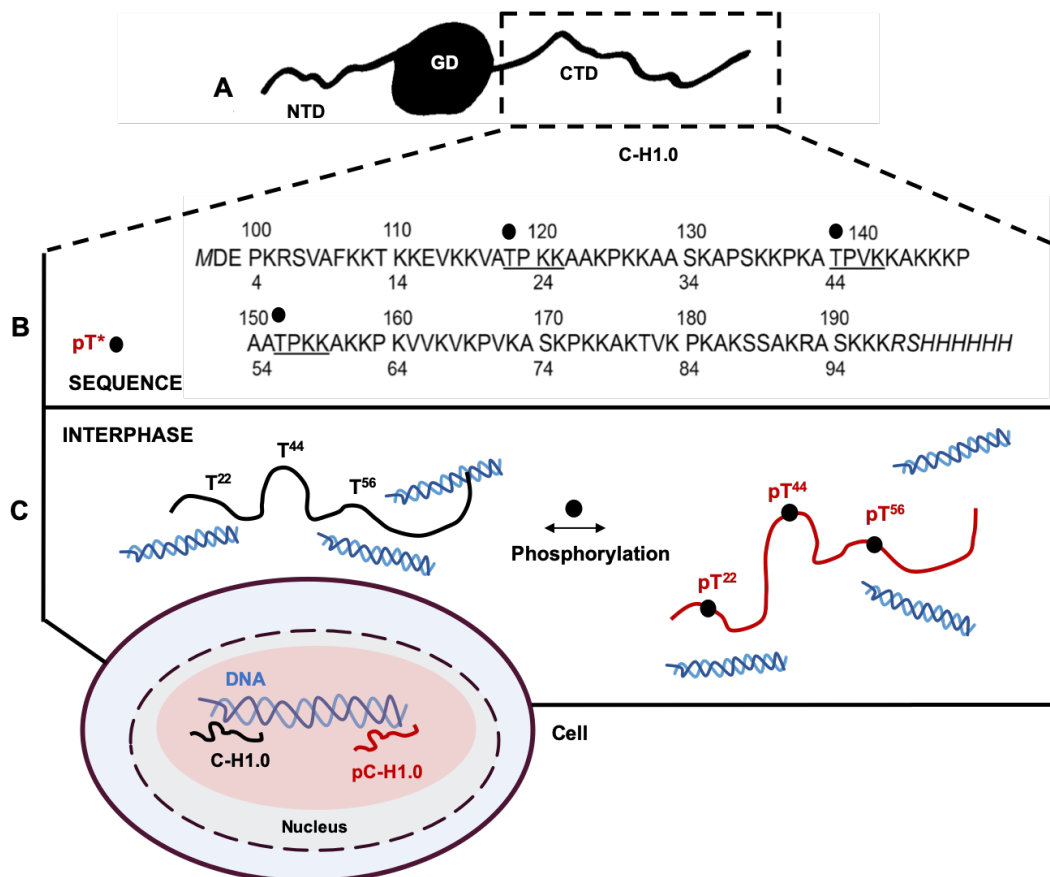

Figure 1. (A) Schematic representation of the structural domains of the Histone H1 molecule. **(B)** Sequence of C-H1.0. The numbers below correspond to the construct used in this work with phosphorylation sites (black circles). **(C)** Interphase cell cycle showing schematically the phosphorylated (red) and non-phosphorylated (black) C-H1.0 and DNA.

3.3   RESULTS AND DISCUSSION

Site-specific DNA interaction on the C-terminal domain of Histone H1.0 (C-H1.0)

To examine the interaction of C-H1.0 with DNA, we selected a non-palindromic double stranded DNA of 26 base pairs (26bp-dsDNA), formed by alternating homopolymer AT tracks (see Materials and Methods; **Figure S1A**). This sequence mimics SAR DNA (scaffold associated regions) to which C-H1 has been reported to bind preferentially [29].

The C-H1.0/DNA interaction was examined by titrating a ^{15}N,^{13}C-C-H1.0 sample with the 26bp-dsDNA, which, in terms of simplicity, will be denoted as DNA from hereon. Because of the low signal dispersion observed in the 2D ^{1}H-^{15}N-HSQC spectra of C-H1.0 (Chapter 2; [28]), the titration was followed by 2D ^{13}C-detected CON experiments, which show less signal overlap (Chapter 2; [28]). Moreover, to be able of following as many cross-peaks as possible, we used 2D ^{13}C-detected CON "folded" spectra to gain resolution in the ^{15}N dimension. So, the spectral width of 2D CON was reduced a 67 % to achieve three times more resolution (**Figure 2A**) without increasing acquisition time.

Upon DNA titration, the cross-peaks present in the "folded" 2D CON experience changes in chemical shift and in intensity (**Figure 2B-C**), which confirms that C-H1.0 interacts with DNA. That DNA and C-H1.0 are interacting under our experimental conditions is further evidenced by the fact that DNA imino protons signals observed in 1D ^{1}H NMR spectra are also shifted relative to their positions in free DNA (**Figure S1B**), as a consequence of different chemical environments in free DNA and in the CH1.0/DNA complex. The cross-peaks intensities in the "folded" 2D CON decrease until complete disappearance, being the titration point at C-H1.0/DNA 1:0.35 ratio the point at highest DNA concentration in which all cross-peaks are still detectable (**Figure 2B-C**). To each titration point (see Materials & Methods), the cross-peaks observed at the "folded" 2D CON were assigned using that of the initial spectrum acquired for the free C-H1.0. When necessary, assignment was confirmed using 3D ^{1}HN-detected HNCO spectra, which were recorded using non-uniform-sampling (NUS) to reduce acquisition time. Interestingly, the chemical shift and intensity changes do not occur evenly for all the cross-peaks (**Figure 2C**).
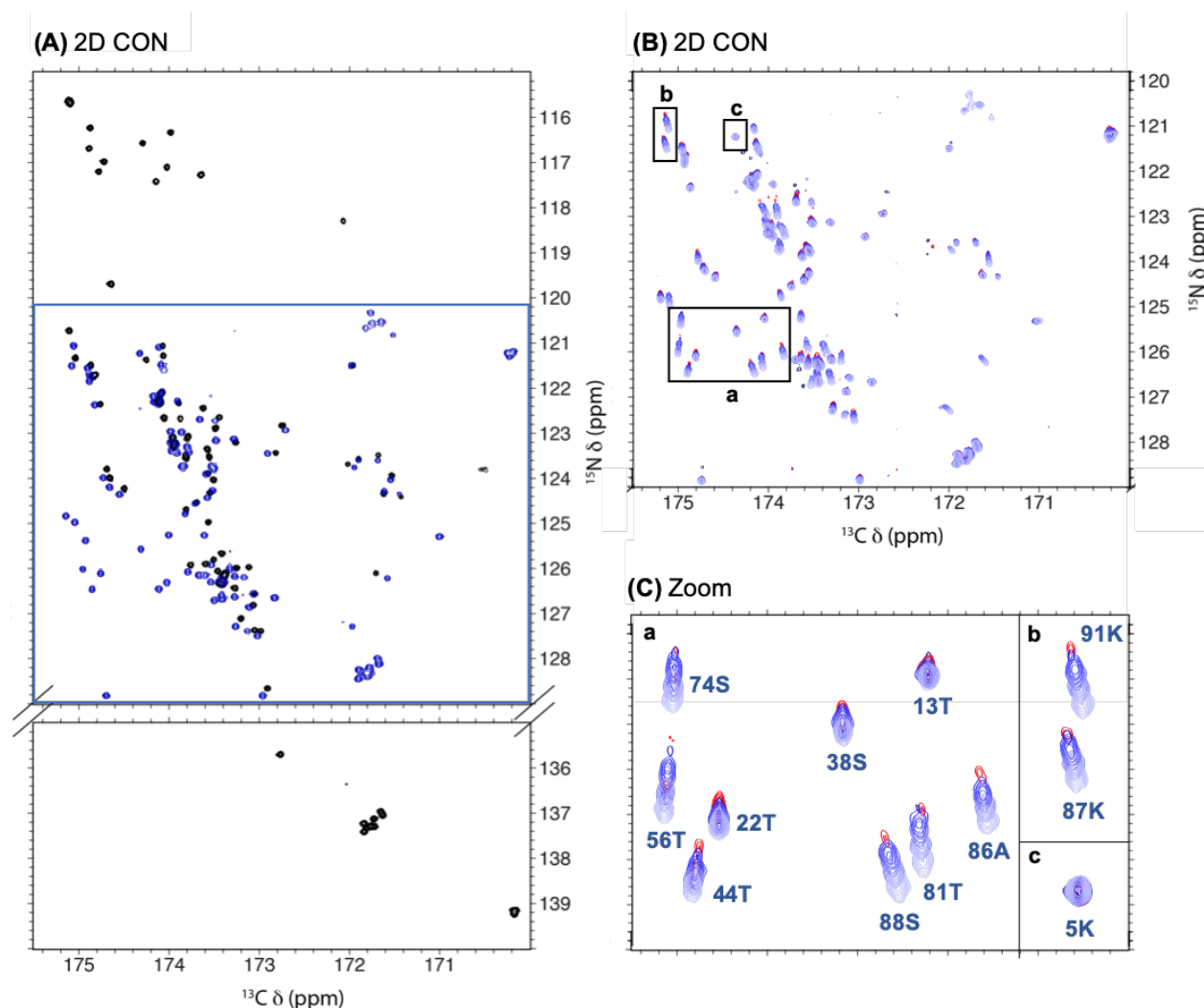
Figure 2. (A) 2D CON spectra of 1 mM ^{15}N, ^{13}C C-H1.0 (black contours) and 2D CON "folded" experiment of 600 μM ^{15}N, ^{13}C C-H1.0 (blue contours). **(B)** 2D CON "folded" spectra of at different [C-H1.0 : DNA] ratios (Material and Methods). The [1 : 0.35] titration point is represented in red color. **(C)** Crowd regions of titration (a, b and c) are zoomed for better visualization. Notice that each cross-peak corresponds to ^{15}N of residue i, and ^{13}C' of residue i − 1.

Hence, we analysed the sequence dependence of: (i) cross-peaks intensity ratios, and (ii) Chemical Shift Perturbations (CSP, see Materials and Methods) (**Figure 3**). The analysis of intensity ratios and CSPs was performed using the data at the C-H1.0/DNA 1:0.35 point, because many cross-peaks become undetectable at higher DNA concentrations (1:0.5 and 1:0.7 points).

Looking at the plots shown in **Figure 3,** it is interesting to note that intensity ratios are smaller, that is, the decrease in cross-peak intensity is larger at the C-terminal region than at the N-terminal. Concerning CSPs, they are smaller at the N-terminal region than at the C-terminal. Thus, the two parameters show that DNA binding affects more the C-terminal region than the N-terminal region, which indicates that C-H1.0 interacts with DNA throughout its C-terminal region. Hence, the C-H1.0/DNA interaction is site-specific. DNA recognition might be due to electrostatic effects, such as the interaction between protein positive charges and DNA phosphates. However, the distribution of the 40 Lys residues is quite homogeneous along the C-H1.0 sequence (**Figure 3B**) showing no correlation with binding specificity
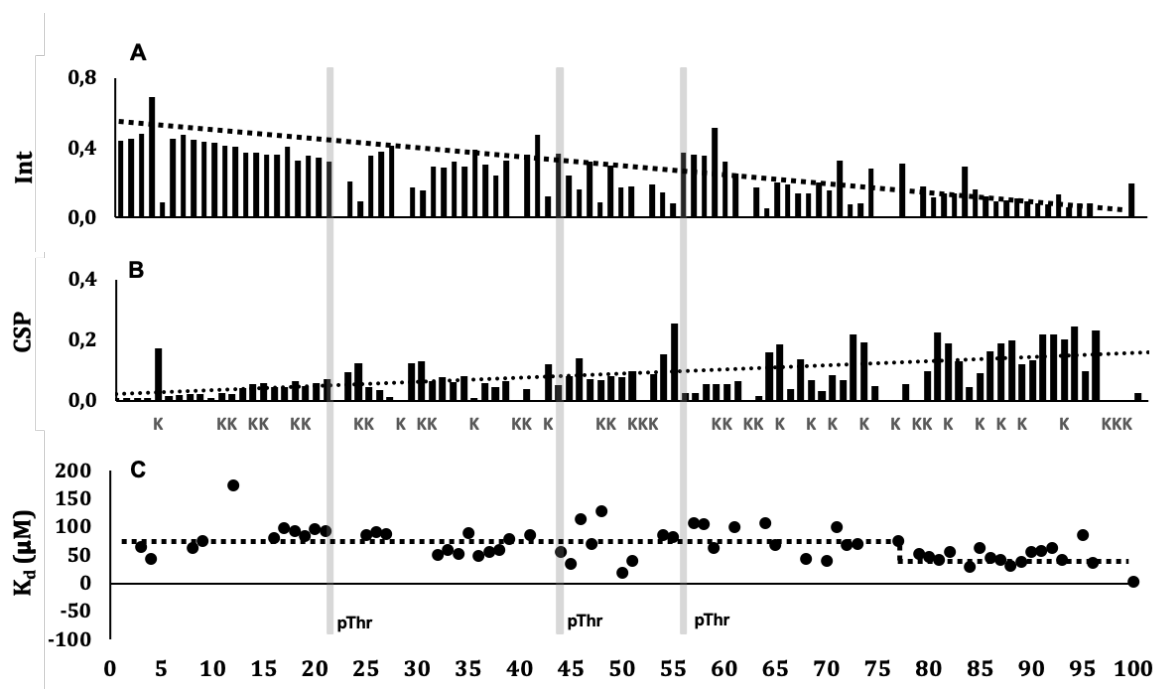
Figure 3. (A) Intensity ratio values (Int) between [C-H1.0 : DNA] at [1 : 0] and [1 : 0.35] titration points of 2D CON "folded" experiments as a function of residue number. **(B)** Chemical Shifts Perturbation (CSP, ppm) in the same titration points that used before. The Lys residues present in C-H1.0 sequence are shown at the X axis. **(C)** K_d values using 2D CON "folded" spectra (black circles). In the three panels, the dotted line shows the tendency, and the grey vertical bars indicate the positions of the Thr residues of the phosphorylation sites. All parameters was measured in 10 mM phosphate buffer at pH 5.5, 10 mM NaCl and 10% D_2O.

Since affinity constants (K_d) can be obtained from the dependence of CSP values on DNA concentration [30], it would be interesting to check whether the K_d values will also reflect a difference between N- and C-terminal regions of C-H1.0. As seen in equation 1 (Materials and Methods), it is necessary to know the C-H1.0 concentration to get the K_d values. This is problematic in the case of C-H1.0, whose only aromatic residue is a Phe residue (Figure 1). Therefore, concentration has to be estimated from peptide bond absorbance at 205 nm (A_{205}; see Materials and Methods), which is less accurate than the usual method of concentration determination applied in proteins containing Tyr and Trp residues, which is based on the absorbance at 280 nm. We corroborated the C-H1.0 concentration derived from A_{205} by comparing signal-to-noise in 1D ^1H NMR spectra of C-H1.0 sample with that in an Ubiquitin reference sample of known concentration. Thus, the K_d values are not accurate, but they are valid for a "per residue" comparison within the C-H1.0 domain. These K_d values determined from CSP $^{(15N,C')}$ (^{15}N and C' chemical shifts measured in "folded" 2D CON spectra) are plotted as a function of C-H1.0 residue number in **Figure 3C**. In concordance with conclusions from the previous analysis based on cross-peaks intensities and CSPs, the residues at the C-terminal region showed smaller K_d values than those at the N-terminal region. Thus, the average K_d obtained for the N-terminal region (residues 1-79) is 77 ± 28 μM and for the C-terminal residues (80 – 100) 49 ± 14 μM. The overall K_d obtained from the average for all residues is 70 ± 28 μM, which is higher than those recently reported for other C-terminal domains of histone, which are in the nM range [31–33]. We do not think that the errors on the measurement of C-H1.0 concentration suffice to explain this difference. However, there are several other explanations, such as the fact that the proteins have some sequence differences, the DNA lengths are different, and the experimental conditions are not the same. In any case, we consider more important to pinpoint the C-H1.0 regions that interact strongly with DNA, than to get an accurate K_d value.

Next, we took a deeper look at the three phosphorylation motifs present in C-H1.0 to see whether there exists any different behaviour among them. **Figure 4** shows the dependence of the CSP and CON cross-peak intensities as a function of DNA concentration for the Thr residues at the motifs (T^{22}, T^{44}, and T^{56}); as well as for E^3 and K^{91} that have been taken as control for the N- and C-terminal regions, respectively.

The curves for these two last residues very clearly show that the N-terminal residue is less affected than the C-terminal residue upon DNA binding. For instance, the residue K[91] experienced the greatest variance in CSP values and the sharpest falls in intensity upon DNA interaction. Concerning the Thr residues, they show an intermediate behaviour, which seems to correlate with their proximity to the N- and C- regions. Thus, T[22], the closest to the N-end, is less affected by DNA interaction than T[44], and T[44] more than T[56], the closest to the C-terminal region.
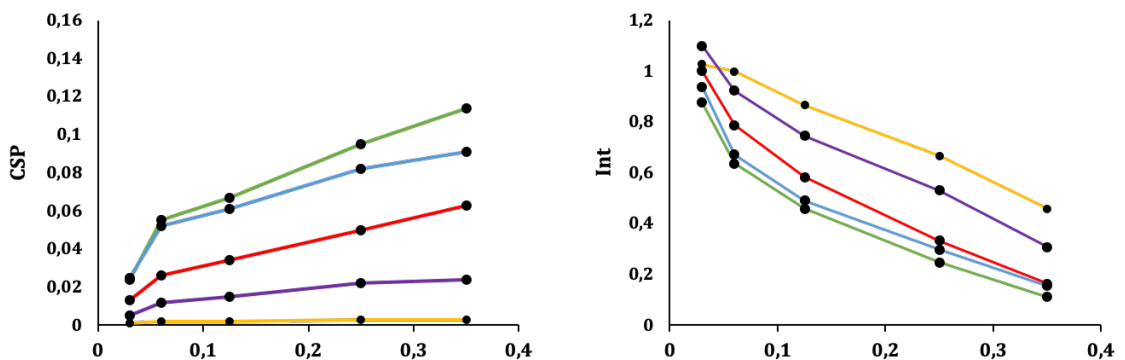
Figure 4. Chemical shift perturbation (CSP, ppm) and normalized intensity (Int) of E[3] (yellow), T[22] (purple), T[44] (red), T[56] (green) and K[91] (blue) as a function of DNA titration points using 2D ^{1}H-^{15}N HSQC experiments (X axis). All parameters were measured in 10 mM phosphate buffer at pH 5.5, 10 mM NaCl and 10 % D_2O.

Non-phosphorylated C-H1.0 is highly disordered, both free and bound to DNA

Since C-H1.0 is an intrinsically disordered domain (see Chapter 2; [28]), it is important to know whether, as occurs in some IDPs (see Introduction), becomes structured or at least more ordered upon DNA-binding. To that aim, we proceeded to examine whether any region show a tendency towards some secondary structure (α-helical or β-sheet) and check its dynamics behavior on fast timescales (subnanoseconds). To delineate secondary structure we used the deviations of the $^{13}C_\alpha$ and $^{13}C'$ chemical shifts from random coil values at the C-H1.0/DNA 1:0.35 point. The $^{13}C_\alpha$ chemical shifts were obtained from a 3D HNCA experiment acquired using NUS (see Materials and Methods). As seen in **Figure 5**, the $\Delta\delta_{C\alpha}$ and $\Delta\delta_{C'}$ shown by DNA-bound C-H1.0 are within the random coil ranges ($|\Delta\delta_{C'}| \leq 0.4$ ppm; $|\Delta\delta_{C\alpha}| \leq 0.4$ ppm), which indicates the absence of any secondary structure tendency. Considering the similitude between the profiles of DNA-bound C-H1.0 to those of free protein, it seems that DNA does not induce any secondary structure formation in C-H1.0. The

disagreement of this result with reported FT-IR data [6–9], which suggested an increase on the content of secondary structure, is probably a consequence of differences in the experimental conditions.

It is also interesting to investigate if the C-H1.0 flexibility is affected by the DNA interaction. Unfortunately the experimental measurement of relaxation parameters was impeded by the loss of signal intensity. Therefore, we decided to get the S^2 order parameter predicted from chemical shifts ($^1H^N$, ^{15}N, $^{13}C\alpha$ and $^{13}C'$) using the TALOS+ program [34]. The resulting averaged S^2 values were 0.36 ± 0.14 for free C-H1.0, and 0.48 ± 0.17 for DNA-bound C-H1.0 using the chemical shifts at the C-H1.0/DNA 1:0.35 point.
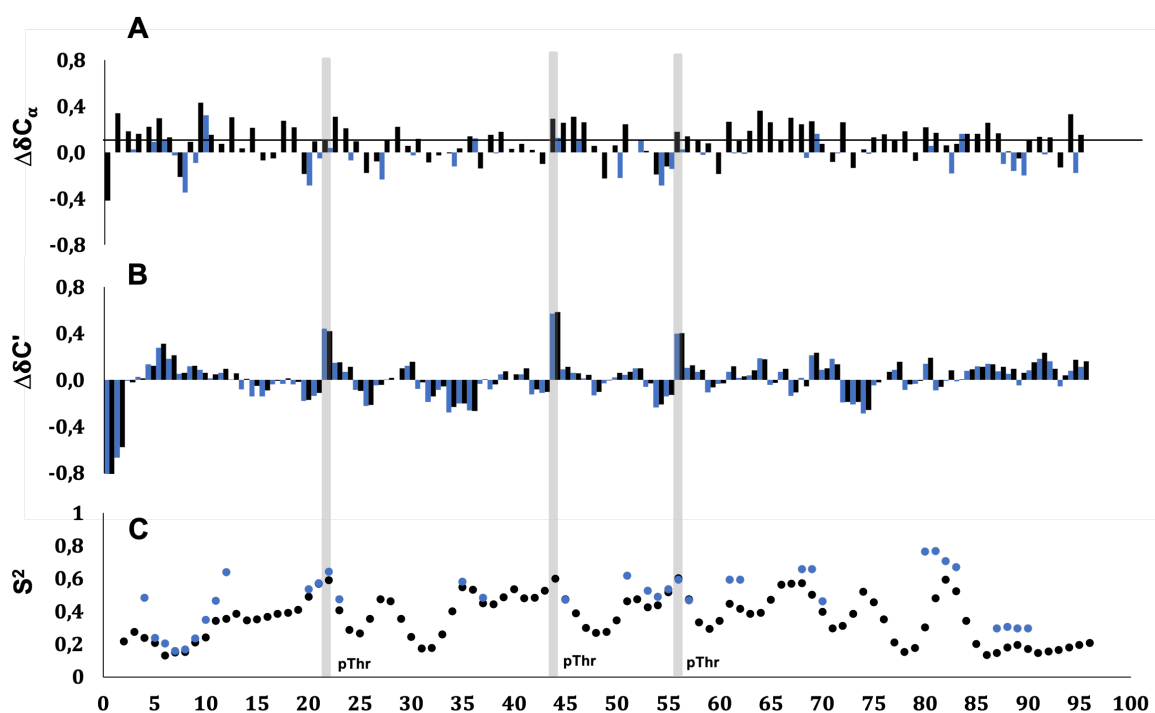
Figure 5. (**A**) $\Delta\delta_{C\alpha}$ values ($\Delta\delta_{C\alpha} = \Delta\delta_{C\alpha}^{observed} - \Delta\delta_{C\alpha}^{reference}$, ppm) from 3D HNCA spectra. (**B**) $\Delta\delta_{C'}$ values ($\Delta\delta_{C'} = \Delta\delta_{C'}^{observed} - \Delta\delta_{C'}^{reference}$, ppm) (panel below) from 2D CON "folded" spectra. (**C**) S2 values predicted from chemical shifts using the TALOS+ program [34]. Free non-phosphorylated C-H1.0 is represented in black and the C-H1.0/DNA 1 : 0.35 point in blue. Grey vertical bars indicate the positions of the phosphorylated sites in Thr residues.

These values are quite close, which points out that the DNA interaction induces a very slight increase in rigidity, if any. Nevertheless, the S^2 values of residues at the C-terminal region are higher at the DNA-bound C-H1.0 than at free C-H1.0. This suggests a decrease in flexibility just at the region, which interacts more strongly with DNA (see previous section).

Effect of phosphorylation on DNA recognition by C-H1.0

To examine the effect of phosphorylation on DNA interaction we used a sample containing equal amounts of non-phosphorylated C-H1.0 and tri-phosphorylated (pT-C-H1.0). We employed this sample to minimise errors due to the estimation of protein concentration, which cannot be accurately measured for C-H1.0 and pT-C-H1.0 because their sequences do not contain Tyr and/or Trp aromatic residues (see above). However, their relative concentrations in a sample containing both proteins can be confirmed from the relative intensities of equivalent cross-peaks in a 2D spectrum. On the other hand, having phosphorylated and non-phosphorylated species in the same sample can mimic the cell conditions at the Interphase, in which both forms coexist, as mention in the Introduction.

An inconvenient of using a sample containing C-H1.0 and pT-C-H1.0 is a loss of information due to overlapping between equivalent cross-peaks of the two species. This is so because 2D NMR spectra (**Figure 6A** and **Figure S2**) contain the cross-peaks of the two proteins, so ideally we would observe two cross-peaks per residue, one for C-H1.0 and another for pT-C-H1.0. Unfortunately, many of them overlap or are too close to get unambiguous data about them even in the 2D CON spectra (**Figure S2**). Since in the case of the "C-H1.0 + pT-C-H1.0" sample, the 2D CON spectra does not solve the ambiguity problem, DNA titration was followed using 2D ^{1}H,^{15}N-HSQC spectra, which are the most sensitive of the two. Moreover, it is feasible to get data about the phosphorylation motifs from the 2D ^{1}H,^{15}N-HSQC spectra. As in the C-H1.0 sample, all the cross-peaks in 2D CON and 2D ^{1}H,^{15}N-HSQC suffer changes in chemical shifts and in intensity upon DNA interaction (**Figure 6B**). In a qualitative way, it was clear that the losses of cross-peak intensity were smaller for the tri-phosphorylated pT-C-H1.0 than for the non-phosphorylated C-H1.0, as highlighted for residues 21 and 44 (**Figure 6B**). This indicates that phosphorylation does not impede DNA recognition, but DNA affinity is lower for pT-C-H1.0 than for C-H1.0, which is in agreement with previous data from spectroscopic studies [33].
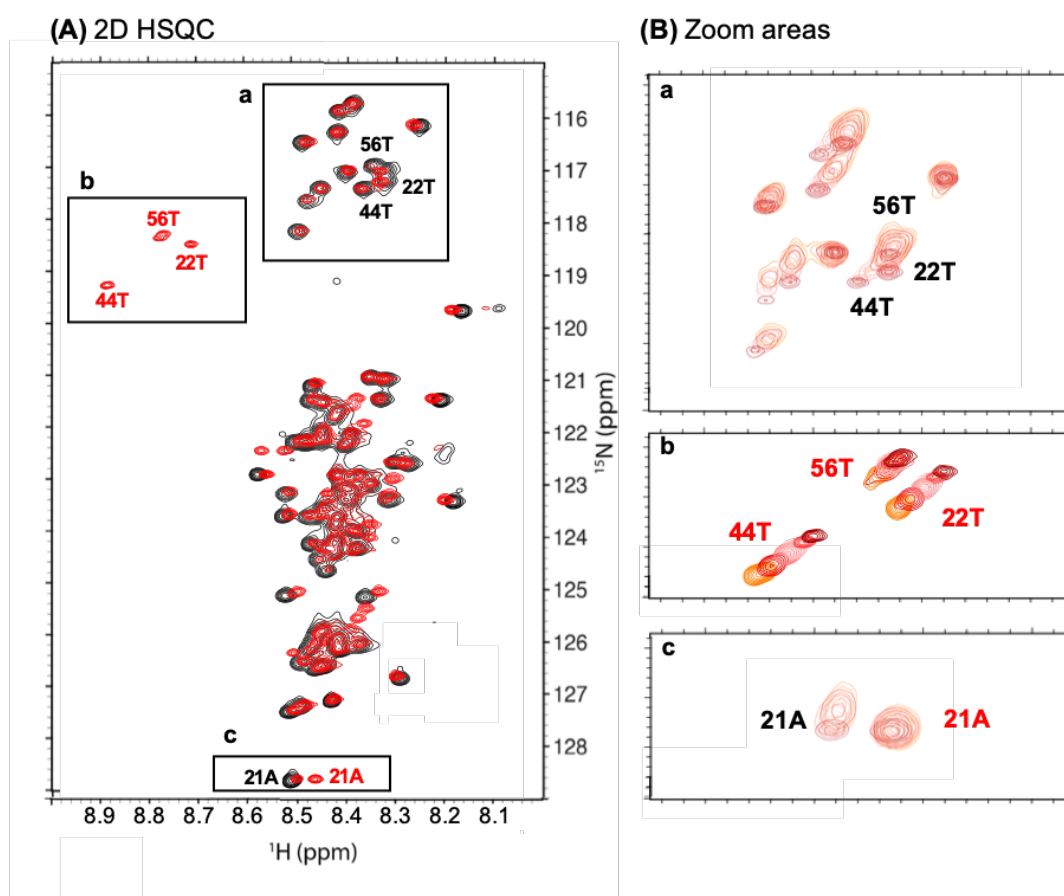
Figure 6. (A) 2D ^{1}H-^{15}N HSQC spectrum of 600 μM ^{15}N, ^{13}C C-H1.0 (black) and of 100 μM ^{15}N, ^{13}C C-H1.0 + 100 μM ^{15}N, ^{13}C pT-C-H1.0 (red). **(B)** Crowd regions of DNA titration (a, b and c) are zoomed for better visualization. The [1 : 0.35] titration point is represented in orange color. The phosphorylated residues are in red color, the non-phosphorylated ones of the 100 μM ^{15}N, ^{13}C C-H1.0 + 100 μM ^{15}N sample are in black.

Figure 7 shows the dependence of the intensities and CSP values with the DNA/protein ratio for A^{21} (pT^{22}-preceding residue from pT-C-H1.0), pT^{44} and pT^{56}. The fact that A^{21}, at the N-terminal region, is less affected (less intensity decrease and smaller CSP) than pT^{56}, closer to the C-terminal region, suggests that phosphorylated pT-C-H1.0 behaves as non-phosphorylated C-H1.0 in that the N-terminal region feels less the presence of DNA than the C-terminal. However, the data for pT^{44} are a bit striking. The changes in intensity would agree with pT^{44} behaving intermediately between A^{21} and pT^{56}, as found for T^{44} in the sample of non-phosphorylated C-H1.0 (**Figure 4**).

However, the CSP dependence of pT^{44} (**Figure 7**) seems to suggest that DNA interaction around this residue (pT^{44}PVK) would be stronger than for A^{21} (as a probe of pT^{22}PKK motif) and pT^{56} (pT^{56}PKK

motif), and even stronger than for the equivalent T[44] in non-phosphorylated C-H1.0 in the "C-H1.0 + pT-C-H1.0" sample (**Figure 7**). A plausible explanation for this apparent contradiction can be found by taking into account that CSP values monitor DNA interaction, but other effects can affect them because of the high sensitivity of chemical shifts to the molecular environment, whereas loss of cross-peak intensity should only be due to DNA interaction.
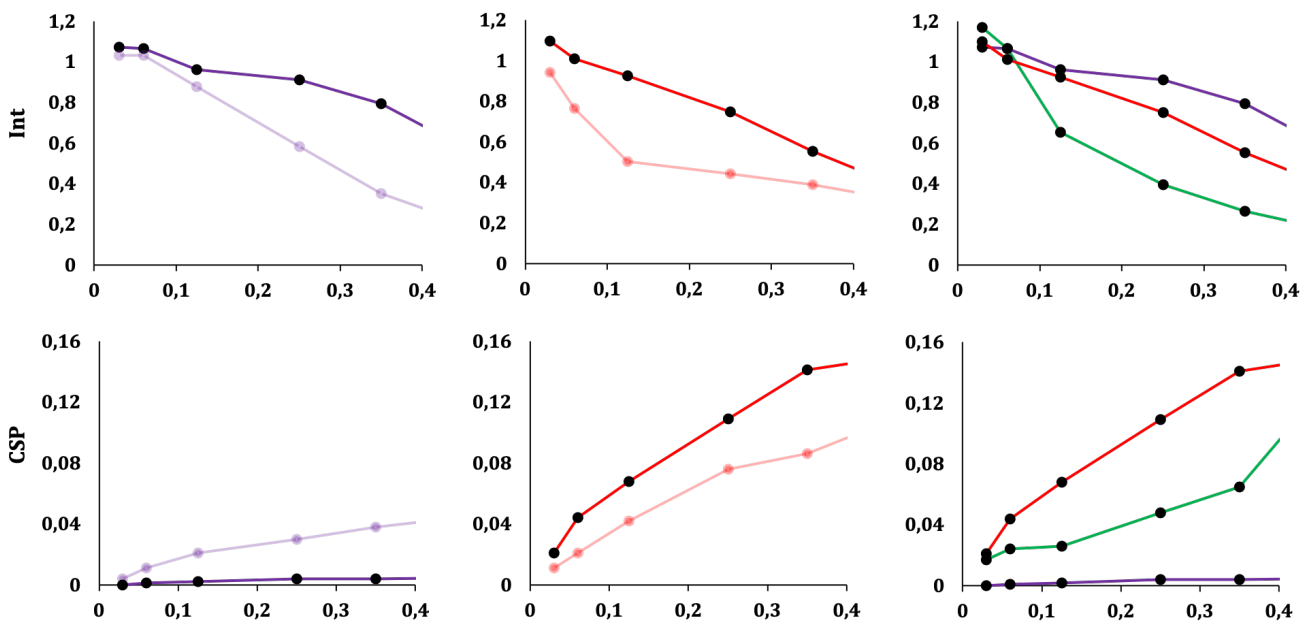
Figure 7. Chemical shift perturbation (CSP, ppm) and normalized intensity (Int) for A[21] (purple), T[44] (red), T[56] (green) as a function of DNA titration points using 2D ^{1}H-^{15}N HSQC experiments acquired in 10 mM phosphate buffer at pH 5.5, 10 mM NaCl and 10% D_2O, and 25°C. Black circles are for C-H1.0 and colour circles for pT-C-H1.0.

Then, it is possible that the conformation of the pT[44]PVK changes upon DNA interaction more than those of the pT[22]PKK and pT[56]PKK motifs, and therefore the changes in chemical shifts (reflected in the CSP values) are larger. We can speculate that this different behaviour is due to (i) an effect of the motif sequence (pTPVK versus pTPKK), which would parallel that the two motifs show different interactions in the conformations of free model peptides (Chapter 4), and (ii) an effect of the repulsions between the negative charges of DNA and pThr phosphates, being the middle motif the most affected.

3.4 CONCLUSIONS

The C-terminal domain of the subtype Histone H1.0 (C-H1.0) plays a key role in gene expression regulation through DNA interaction, which, in turn, is regulated by phosphorylation at the Thr residue present in the ^{118}TPKK^{121}, ^{140}TPVK^{143} and ^{152}TPKK^{155} motifs. To get structural details at residue level into the C-H1.0 / DNA interaction we titrated a ^{13}C,^{15}N-C-H1.0 sample with a non-palindromic double stranded DNA of 26 base pairs, which mimics the target SAR DNA of the C-terminal domain of Histone H1. Upon DNA titration, which was followed using 2D ^{13}C-detected CON spectra, and not only ^{1}H,^{15}N-HSQC, because of its better signal dispersion, the cross-peaks in these 2D spectra are shifted and decrease in intensity.

The non-uniform distribution of these changes indicates that DNA interaction occurs preferentially at the C-terminal region, which might become slightly less flexible. Concerning the phosphorylation motifs, their DNA affinity increases with their closeness to the C-terminal region. According to C' and ^{13}Cα chemical shift deviations, no secondary structure is triggered by DNA binding. This is in contrast to previous spectroscopic data suggesting that DNA induces some secondary structure in C-H1.0, but this difference might be a consequence of the sensitivity of protein/DNA interactions to experimental conditions. But, it is in agreement with results in the C-terminal domain of a different Histone H1, which remains disordered in its DNA-bound state [33]. Thus, our results indicates that some IDPs do not become ordered upon binding its target partner.

As concerns the effect of phosphorylation, DNA interaction is weaker with tri-phosphorylated pT-C-H1.0 than with the non-phosphorylated C-H1.0, as previously reported [33], but very probably it also occurs by the C-terminal region. However, the behavior of the middle phosphorylation motif, pT^{138}PVK, upon DNA binding differs from those of the pT^{118}PKK and pT^{152}PKK motifs, and also from that of T^{138} in C-H1.0. This can be attributed to an effect of the motif sequence or of the position of the motif in the sequence. The local change in the conformational ensemble around the middle motif between the free and DNA-bound pT-C-H1.0, together with the weaker DNA-affinity of phosphorylated pT-C-H1.0 relative to non-phosphorylated C-H1.0 might account for the role played by Histone phosphorylation in regulating chromatin condensation/de-condensation along the cell cycle.

3.5   MATERIALS AND METHODS

Expression and purification of C-H1.0

The sequence of the C-H1.0 was cloned in the pQE-60 vector (Qiagen), as previously described [29]. The recombinant expression vector was transformed into *E. coli* BL21 (DE3) competent cells expressing in kanamycin (30 μg/l). The isotopically labeled media used was K-MOPS supplemented with ^{13}C-D-glucose (Cambridge Isotope Laboratories Inc.) and ^{15}NH$_4$Cl (Cambridge Isotope Laboratories Inc.) [35]. Briefly, cells were grown to an OD$_{600nm}$ of 0.6-0.8 and the protein expression was induced with 1 mM IPTG, allowing this expression to proceed for 24 h at 25 °C. Next, cells were lysed in the lysis buffer (0.05 M NaH$_2$PO$_4$, 0.75 M NaCl, 0.02 M imidazol and 4 M guanidine hydrochloride, pH 8.0) for 15 min at 4 ºC. The extract was sonicate (3 min ON, 10 min OFF and 30 % Amplitude) and centrifuged at 15000 rpm for 15 min. The supernatant was loaded on a HiTrap chelating HP column (GE Healthcare) equilibrated with lysis buffer.

Then, the column was washed using 40 mM imidazole, and finally, the protein was eluted with 250 mM imidazol in lysis buffer and desalted by gel filtration through Sephadex G-25 (Sigma-Aldrich). All process was done at 4 ºC and the protein was storage at -20 ºC.

Preparation of the C-H1.0 + pT-C-H1.0 sample

The isotopically labeled C-H1.0 was phosphorylated *in vitro* using CDK2-cyclin A2 kinase (Sigma Aldrich) as previously described [25]. The 100% phosphorylation reaction requires an specific buffer (50 mM Tris-HCl, 10 mM MgCl$_2$,1 mM EGTA, 20 mM dithiothreitol (DTT) and 200 μM ATP, at pH 7.5) and 1U of CDK2-cyclin A per 5 μg of C-H1.0. We put 10 μg of C-H1.0 to obtain a sample 50% C-H1.0 and 50% pT-C-H1.0. After a hour of incubation at 30 ºC, the reaction buffer was eliminated using HiTrap desalting column (GE Healthcare).

Preparation of double stranded DNA (dsDNA)

Single stranded 26-mer oligonucleotides (fwDNA and bwDNA) were purchased from Macrogen Humanizing genomics (Madrid, Spain).

> fwDNA (5' GAT ATT TAT ATT TAT ATT TAT ATT TG 3')
>
> MW calculated = 7969.2; MW measured = 7997.9; GC% = 7.69
>
> bwDNA (5' CAA ATA TAA ATA TAA ATA TAA ATA TC 3')
>
> MW calculated = 7961.2; MW measured = 7996.7; GC% = 7.69

Oligonucleotides fwDNA and bwDNA were dissolved in 10 mM phosphate at pH 7.5 and 10 mM NaCl annealing buffer at 5 mM concentration. The two strands were combined in equimolar amounts and annealed using the G-Storm GS1 thermal cycler. The annealing protocol consisted of: (i) heating up to 90° C; (ii) remain at 90° C for 2 min; and (iii) a ramp to lower the temperature from 90° C to 10° C at 0.2° C/cycle. The resulting 26-base pairs dsDNA at 2.5 mM concentration was stored at 4° C.

To confirm that the dsDNA was formed upon the annealing protocol, we acquired 1D ^{1}H-NMR spectra of dsDNA, fwDNA and bwDNA in aqueous solution (**Figure S1A**). The first indication about dsDNA formation comes from the fact that the spectrum of dsDNA does not correspond to the sum of the spectra of the single stranded fwDNA and bwDNA. Stronger evidence about dsDNA being indeed double-stranded is provided by the presence of imino protons in the 12.0-15.0 ppm region of the NMR spectrum (**Figure S1A**), which are characteristic for double-stranded DNAs. No signal was observed at that region in the NMR spectra of the single-stranded fwDNA and bwDNA.

Measurement of protein concentration

Protein concentration was derived from the absorbance at 205 nm (A^{205}), as described [36] and corroborated from comparing signal-to-noise at the methyl region in 1D ^{1}H NMR spectra of C-H1.0 sample with that in an Ubiquitin reference sample of known concentration.

DNA titrations' samples

The following samples were used for the titration experiments:

> **(i)** ^{15}N,^{13}C-C-H1.0 sample: 600 µM in 10 mM phosphate buffer at pH 5.5, 10 mM NaCl and 10% D$_2$O, containing DSS as internal reference.

(ii) 50% ^{15}N,^{13}C-pT-C-H1.0 + 50% ^{15}N,^{13}C-CH1.0 sample: 100 μM ^{15}N,^{13}C-pT-C-H1.0 + 100 μM ^{15}N,^{13}C- C-H1.0 in 10 mM phosphate buffer at pH 5.5, 10 mM NaCl and 10% D$_2$O.

(iii) Stock dsDNA solution: 2.5 mM in 10 mM phosphate buffer at pH 7.5 and 10 mM NaCl

Table 1 lists the DNA concentrations and the protein/DNA ratios at all the titrations points.

	C-H1.0 titration		C-H1.0 + pT-H1.0 titration	
Table 1. DNA concentration and protein / DNA ratios at the different titration points				
	DNA conc.	C-H1.0/DNA ratio	DNA conc.	(C-H1.0 + pT-C-H1.0) / DNA ratio
Initial point	0	1:0	0	1:0
DNA1	18.65 μM	1:0.03	3.21 μM	1:0.03
DNA2	36.94 μM	1:0.06	6.23 μM	1:0.06
DNA3	60.97 μM	1:0.125	12.44 μM	1:0.125
DNA4	119 μM	1:0.25	24.75 μM	1:0.25
DNA5	163.55 μM	1:0.35	34.52 μM	1:0.35
DNA6	227.27 μM	1:0.5	49.02 μM	1:0.5
DNA7	363.25 μM	1:0.7	96.15 μM	1:0.7

NMR experiments

All NMR experiments were collected at 298 K on a Bruker Avance spectrometer, operating at a ^{1}H frequency of 800.1 MHz equipped with a TCI cryoprobe. The set of collected experiments for each protein and each titration point consisted of 1D ^{1}H spectrum, 2D CON ^{13}C-detected experiment and two ^{1}H-detected experiments, i.e., 2D ^{1}H–^{15}N HSQC and 3D HNCO. Furthermore, a 3D HNCA experiment was run for C-H1.0 to confirm secondary structure. In the case of 3D spectra, the less sensitive ^{13}C-detected experiments were recorded using linear sampling, and the most sensitive ^{1}H-detected using non-uniform sampling (NUS).

The 1D ^{1}H NMR experiment with ^{15}N decoupling during acquisition was acquired with 2048 complex data points with the carrier set at 4.75 ppm, spectral width of 16 ppm and 16 scans. The total measurement time was 21 seconds. The 2D ^{13}C-detected CON folded experiment was acquired with 512 and 64 complex points in the ^{13}C' direct dimension and ^{15}N indirect dimension. The carriers of ^{13}C' and ^{15}N dimensions were set at 172.5 ppm and 124.7 ppm, respectively, and the spectral widths were 10 and 9 ppm in the ^{13}C' and ^{15}N dimensions, respectively. The total measurement time was 1 h and 40 min. For the 2D ^{1}H–^{15}N HSQC spectrum 1K and 128 complex data points were acquired for direct and indirect dimension, respectively, two scans were accumulated and the total experimental time was 10 min. The 3D HNCO and HNCA experiments were acquired with non-uniform sampling (NUS). They were acquired using a 204-complex point sampling schedule with 8 scans per FID in 1h and 8min each one. The maximum increment in the NUS schedule is 32 for the ^{15}N attached directly to ^{1}H acquired in the direct dimension (f2) and ^{13}C (f1) dimensions. ^{15}N (f2) is centred in 122 ppm with the spectral width being 1297.33 Hz, and ^{3}C (f1) are centred in 172.5 ppm and 54 ppm with the spectral width being 1297.33 Hz and 3880.5 Hz, respectively. The directly observed ^{1}H dimension has a spectral width of 8802.817 Hz (centred at 4.75 ppm) with 1024 complex points.

For all the experiments, the resulting matrix was zero filled to double the number of original points in all dimensions and shifted squared sine-bell apodization functions were applied in all dimensions prior to Fourier transformation. Spectra were processed using either TOPSPIN v2.1 pl6 (Bruker, Inc) or NMRpipe [37] and the istHMS reconstruction method was used to process NUS data [38]. Finally, they were analyzed with the programs SPARKY (T. Goddard and D.G. Kneller, SPARKY 3, University of California, San Francisco, USA) and/or NMRview [39].

Chemical shift perturbation (CSP) from ^{1}H-^{15}N HSQC (X= ^{1}H; Y = ^{15}N) or CON (X= ^{13}C; Y = ^{15}N) experiments were obtained by applying the following equation: CSP = $[(\delta_X^{pT-C-H1.0} - \delta_X^{C-H1.0})^2 + ((\delta_Y^{pT-C-H1.0} - \delta_Y^{C-H1.0}) A)^2]^{1/2}$ where the scaling factor A, which is the ratio between X and Y spectral widths (A = SW_X/SW_Y), is equal to 0.4 and 1.1, respectively [40]. CSP values also have been used to measure ligand affinity (K_d).

95

We obtained these results attending to the following equation (where R is the DNA to C-H1.0 ratio) [30]:

Equation 1.
$$CSP = CSP_{max} \ \frac{R}{\frac{K_d}{[C-H1.0]}+R}$$

Cross-peaks intensities in ^{1}H-^{15}N HSQC and CON spectra were normalized relative to the first titration point to examine their dependence with DNA concentration. Values above 1 (Figures 4 and 7) are due to experimental error.

3.6 ACKNOWLEDGMENTS

3.7   REFERENCES

1.   Bednar J, Hamiche A, Dimitrov S. H1-nucleosome interactions and their functional implications. Biochim Biophys Acta - Gene Regul Mech. 2016;1859(3):436–43. Available from: http://dx.doi.org/10.1016/j.bbagrm.2015.10.012

2.   Thoma F, Koller T, Klug A. Involvement of histone H1 in the organization of t... [J Cell Biol. 1979] - PubMed result. J Cell Biol. 1979;83(2 Pt 1):403–27. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2111545&tool=pmcentrez&rendertype=abstract

3.   Wolffe AP, Khochbin S, Dimitrov S. Alan P. Wolffe, Saadi Khochbin and Stefan Dimitrov. BioEssays. 1997;19(3):249–55.

4.   Roque A, Ponte I, Suau P. Interplay between histone H1 structure and function. Biochim Biophys Acta . 2015/09/30. 2016;1859(3):444–54. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26415976

5.   HARTMAN PG, CHAPMAN GE, MOSS T, BRADBURY EM. Studies on the Role and Mode of Operation of the Very-Lysine-Rich Histone H1 in Eukaryote Chromatin: The Three Structural Regions of the Histone H1 Molecule. Eur J Biochem. 1977;77(1):45–51.

6.   Vila R, Ponte I, Collado M, Arrondo JL, Suau P. Induction of secondary structure in a COOH-terminal peptide of histone H1 by interaction with the DNA: an infrared spectroscopy study. J Biol Chem . 2001/06/20. 2001;276(33):30898–903. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11413144

7.   Vila R, Ponte I, Collado M, Arrondo JL, Jimenez MA, Rico M, et al. DNA-induced alpha-helical structure in the NH2-terminal domain of histone H1. J Biol Chem. 2001/10/05. 2001;276(49):46429–35. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11584004

8.   Roque A, Iloro I, Ponte I, Arrondo JL, Suau P. DNA-induced secondary structure of the carboxyl-terminal domain of histone H1. J Biol Chem. 2005/07/12. 2005;280(37):32141–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16006555

9.   Roque A, Ponte I, Suau P. Post-translational modifications of the intrinsically disordered terminal domains of histone H1: effects on secondary structure and chromatin dynamics. Chromosoma. 2016/04/22. 2017;126(1):83–91. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27098855

10.   Ponte I, Vidal-Taboada JM, Suau P. Evolution of the vertebrate H1 histone class: Evidence for the functional differentiation of the subtypes. Mol Biol Evol. 1998;15(6):702–8.

11.   Ponte I, Romero D, Yero D, Suau P, Roque A. Complex Evolutionary History of the Mammalian Histone H1.1-H1.5 Gene Family. Mol Biol Evol. 2017/01/20. 2017;34(3):545–58. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28100789

12.   Millán-Ariño L, Izquierdo-Bouldstridge A, Jordan A. Specificities and genomic distribution of somatic mammalian histone H1 subtypes. Biochim Biophys Acta - Gene Regul Mech. 2016;1859(3):510–9. Available from: http://dx.doi.org/10.1016/j.bbagrm.2015.10.013

13. Lu X, Hansen JC. Identification of Specific Functional Subdomains within the Linker Histone H10 C-terminal Domain. J Biol Chem. 2004;279(10):8701–7.

14. Widlak P, Kalinowska M, Parseghian MH, Lu X, Hansen JC, Garrard WT. The histone H1 C-terminal domain binds to the apoptotic nuclease, DNA fragmentation factor (DFF40/CAD) and stimulates DNA cleavage. Biochemistry. 2005/05/25. 2005;44(21):7871–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15910001

15. Contreras A, Hale TK, Stenoien DL, Rosen JM, Mancini MA, Herrera RE. The Dynamic Mobility of Histone H1 Is Regulated by Cyclin/CDK Phosphorylation. Mol Cell Biol. 2003;23(23):8626–36.

16. Hendzel MJ, Lever MA, Crawford E, Th'ng JP. The C-terminal domain is the primary determinant of histone H1 binding to chromatin in vivo. J Biol Chem. 2004/02/27. 2004;279(19):20028–34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14985337

17. Ponte I, Vila R, Suau P. Sequence complexity of histone H1 subtypes. Mol Biol Evol. 2003;20(3):371–80.

18. Galea CA, Wang Y, Sivakolundu SG, Kriwacki RW. Regulation of cell division by intrinsically unstructured proteins: Intrinsic flexibility, modularity, and signaling conduits. Biochemistry. 2008;47(29):7598–609.

19. Wolffe AP. Packaging principle: How DNA methylation and histone acetylation control the transcriptional activity of chromatin. J Exp Zool. 1998;282(12):239–44.

20. Gréen A, Sarg B, Gréen H, Lönn A, Lindner HH, Rundquist I. Histone H1 interphase phosphorylation becomes largely established in G 1or early S phase and differs in G 1between T-lymphoblastoid cells and normal T cells. Epigenetics and Chromatin. 2011;4(1):15. Available from: http://www.epigeneticsandchromatin.com/content/4/1/15

21. Talasz H, Helliger W, Puschendorf B, Lindner H. In vivo phosphorylation of histone H1 variants during the cell cycle. Biochemistry. 1996;35(6):1761–7.

22. Suzuki M, Gerstein M, Johnson T. An nmr study on the dna-binding spkk motif and a model for its interaction with dna. Protein Eng Des Sel. 1993;6(6):565–74.

23. Lopez R, Sarg B, Lindner H, Bartolome S, Ponte I, Suau P, et al. Linker histone partial phosphorylation: effects on secondary structure and chromatin condensation. Nucleic Acids Res . 2015/04/15. 2015;43(9):4463–76. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25870416

24. Liao R, Mizzen CA. Interphase H1 phosphorylation: Regulation and functions in chromatin. Biochim Biophys Acta. 2015/12/15. 2016;1859(3):476–85. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26657617

25. Roque A, Ponte I, Arrondo JL, Suau P. Phosphorylation of the carboxy-terminal domain of histone H1: effects on secondary structure and DNA condensation. Nucleic Acids Res . 2008/07/18. 2008;36(14):4719–26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18632762

26. Roque A, Ponte I, Arrondo JL, Suau P. Phosphorylation of the carboxy-terminal domain of histone H1 : effects on secondary structure and DNA condensation. Nucleic Acids Res. 2008;36(14):4719–26.

27.	Raghuram N, Strickfaden H, McDonald D, Williams K, Fang H, Mizzen C, et al. Pin1 promotes histone H1 dephosphorylation and stabilizes its binding to chromatin. J Cell Biol . 2013/10/09. 2013;203(1):57–71. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24100296

28.	Chaves-Arquero B, Pantoja-Uceda D, Roque A, Ponte I, Suau P, Jiménez MA. A CON-based NMR assignment strategy for pro-rich intrinsically disordered proteins with low signal dispersion: the C-terminal domain of histone H1.0 as a case study. J Biomol NMR. 2018;72(3–4):139–48. Available from: http://dx.doi.org/10.1007/s10858-018-0213-2

29.	Roque A, Orrego M, Ponte I, Suau P. The preferential binding of histone H1 to DNA scaffold-associated regions is determined by its C-terminal domain. Nucleic Acids Res. 2004/11/25. 2004;32(20):6111–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15562002

30.	Juhl DW, van Rensburg W, Bossis X, Vosloo JA, Rautenbach M, Bechinger B. Tyrocidine A interactions with saccharides investigated by CD and NMR spectroscopies. J Pept Sci. 2019;25(5):1–11.

31.	Machha VR, Waddle JR, Turner AL, Wellman S, Le VH, Lewis EA. Calorimetric studies of the interactions of linker histone H10 and its carboxyl (H10-C) and globular (H10-G) domains with calf-thymus DNA. Biophys Chem. 2013;184:22–8.

32.	Machha VR, Jones SB, Waddle JR, Le VH, Wellman S, Lewis EA. Exploring the energetics of histone H1.1 and H1.4 duplex DNA interactions. Biophys Chem. 2014;185(January 2014):32–8.

33.	Turner AL, Watson M, Wilkins OG, Cato L, Travers A, Thomas JO, et al. Highly disordered histone H1–DNA model complexes and their condensates. Proc Natl Acad Sci U S A. 2018;115(47):11964–9.

34.	Shen Y, Delaglio F, Cornilescu G, Bax A. TALOS+: A hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR. 2009;44(4):213–23.

35.	Neidhardt FC, Bloch PL, Smith DF. Culture medium for enterobacteria. J Bacteriol. 1974;119(3):736–47.

36.	Scopes RK. Measurement of protein by spectrophotometry at 205 nm. Anal Biochem . 1974;59(1):277–82. Available from: http://www.sciencedirect.com/science/article/pii/0003269774900347

37.	Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol Nmr. 1995/11/01. 1995;6(3):277–93. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8520220

38.	Hyberts SG, Frueh DP, Arthanari H, Wagner G. FM reconstruction of non-uniformly sampled protein NMR data at higher dimensions and optimization by distillation. J Biomol Nmr. 2009/08/26. 2009;45(3):283–94. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19705283

39.	Johnson BA, Blevins RA. NMR View: A computer program for the visualization and analysis of NMR data. J Biomol Nmr. 1994/09/01. 1994;4(5):603–14. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22911360

40.	Williamson MP. Using chemical shift perturbation to characterise ligand binding. Prog Nucl Magn Reson Spectrosc . 2013/08/22. 2013;73:1–16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23962882

3.8   SUPPORTING INFORMATION

Figure S1. (A) 1D ^{1}H NMR spectra of free dsDNA (blue), fwDNA (light blue) and bwDNA (light purple) in 10 mM phosphate buffer at pH 5.5, 10 mM NaCl and 10% D_2O. **(B)** Imino region of the 1D ^{1}H NMR spectra of free dsDNA (blue) and in the presence of non-phosphorylated C-H1.0 (black) at saturation titration point [Histone : DNA] [1 : 2].
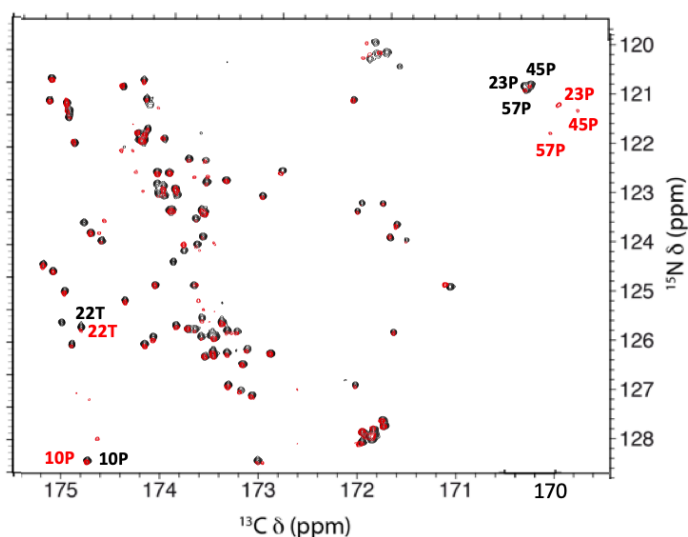
Figure S2. 2D CON "folded" of 600 μM ^{15}N, ^{13}C C-H1.0 (black) and 2D CON "folded" of 100 μM ^{15}N, ^{13}C C-H1.0 + 100 μM ^{15}N, ^{13}C pT-C-H1.0 (red).

CHAPTER 4

Effect of phosphorylation in peptides derived from the intrinsically disordered C-terminal domain of Histone H1.0

4.1. ABSTRACT

H1 linker histones bind DNA regions on the nucleosome surface and play a regulatory role in transcription. DNA binding is modulated by phosphorylation at the intrinsically disordered C-terminal domain. The C-terminal domain of Histone H1.0 subtype (C-H1.0) is phosphorylated in three Thr residues: T118 and T152 in TPKK motifs and T140 in a TPVK motif. The short-life of the full-length C-H1.0, both non-phosphorylated and tri-phosphorylated, complicates NMR studies. Therefore, to get details into the structural consequences of phosphorylation, we followed a minimalist approach consistent in NMR and CD studies of model C-H1.0-derived peptides: T^{118}-H1.0 (residues 103-124) and T^{140}-H1.0 (residues 138-148). Phosphorylation slightly affects the Pro *cis/trans* isomerism; the *cis* percentage decreases in the T^{118}PKK motif and increases in the T^{140}PVK motif. Because of the low *cis* percentages (< 10%), structural behaviour was examined only for the major *trans* species. As in the full-length C-H1.0, the non-phosphorylated and phosphorylated peptides are mainly disordered in aqueous solution, but they become structured in the presence of trifluoroethanol (TFE). The structure of T^{118}-H1.0 contains two almost perpendicular helical regions: a long amphipathic α-helix spanning residues 104-115 and a short $\alpha/3_{10}$ helix, in which T118 or P119 belonging to the phosphorylation motif acts as N-cap. The phosphorylated pT^{118}-H1.0 shows the same two helical segments, but the orientation between them is poorly defined and different from that in non-phosphorylated T^{118}-H1.0. Non-phosphorylated T^{140}-H1.0 and phosphorylated pT^{140}-H1.0 showed very similar α-helices at residues 141-147. The backbone conformation of the TPKK and TPVK motifs is the same in both the non-phosphorylated and phosphorylated states. However, the two motifs differ in side-chain contacts; Thr and pThr side chains interact with the i+2 Lys side chain in the TPKK motif, and with the i+3 Lys side chain in the TPVK motif. These differences might be biologically relevant and be related to different roles for the various phosphorylation motifs present in the C-terminal domain of Histone H1.0. On the whole, the minimalist strategy is validated by the fact that the model peptides behave as the corresponding regions in the full-length C-H1.0, but it provides structural details difficult to get in short-lived intrinsically disordered proteins and domains..

Keywords: Histone, IDP, NMR, peptide model, peptide structure

4.2. INTRODUCTION

H1 linker histones bind to DNA regions on the nucleosome surface, and are thought to be responsible for chromatin condensation and play a regulatory role in transcription [1]. Eukaryotic linker histones are organized in three domains, of which the middle one is globular, and the N- (NTD) and the C-terminal (CTD) ones are intrinsically disordered in their free states [2–4]. Based on FT-IR and CD studies, these NTD and CTD domains were reported to acquire some secondary structure, including turns, α-helices and β-conformations, in the presence of trifluoroethanol (TFE) and upon DNA-binding [2,3].

The multiple subtypes of mammalian H1 histones differ in chromatin affinity, genomic localization, expression pattern and post-translational modifications (PTMs) [5]. H1 CTDs diversity seems to determine their distinct chromatin-binding affinities [6] due to differences in net positive charge and PTMs that decreased it, like acetylation and phosphorylation [4]. This later PTM is done at the consensus (S/T)-P-X-(K/R) motifs by cyclin-dependent kinases (CDKs). In the subtype H1.0, phosphorylation by CDK2 occurs at three Thr residues of the CTD domain, Thr118, Thr140 and Thr152 (**Figure 1**) resulting in a moderate effect on DNA binding affinity and aggregating capacity. Based on FTIR analysis, phosphorylation hardly affects the secondary structure contents of the free domain, but it decreases the α-helix content and increases the β-structure in DNA-bound H1.0 CTD [7]. Phosphorylation-triggered conformational change was also observed in full-length H1 by FTIR in chromatin and by FRET analysis of reconstituted nucleosomes [8,9].

Based on experimental evidences showing that dephosphorylation of two different H1 subtypes was dependent on the *cis-trans* prolyl-isomerase activity Pin1 *in vivo* [8], it could be hypothesized that *cis-trans* proline isomerization plays a role in the observed phosphorylation-induced conformational change in DNA-bound H1 CTD [7].

Most of this structural information comes from spectroscopic techniques, which do not give us details at atomic level. Being the H1 CTD an intrinsically disordered domain, solution NMR is the most appropriate method to get structural data at atomic level. Therefore, we recently focussed on the H1.0 subtype and performed a solution NMR characterisation of its C-terminal domain (C-H1.0; Chapter 2; [10]). Considering the problems found in the study of the full-length domain, in particular,

the short sample life, due to degradation and aggregation, we decided to explore a minimalist approach to gain insights into the structural consequences of phosphorylation on C-H1.0 in a simplest way. To that end, we have designed model peptides for the phosphorylation motifs of CH1.0 (^{118}TPKK^{121} and ^{140}TPVK^{143}; **Figure 1**). Herein, we report NMR and CD studies of two non-phosphorylated and phosphorylated C-H1.0-derived peptides in aqueous solution and in the presence of TFE. Each peptide contains a single phosphorylation motif. We compare the structural behaviour between non-phosphorylated (T^{118}-H1.0 and T^{140}-H1.0) and phosphorylated (pT^{1118}-H1.0 and pT^{140}-H1.0) peptides, and between the isolated peptides and the corresponding regions in the full-length C-H1.0 and pT-C-H1.0 (Chapter 2; [10]).
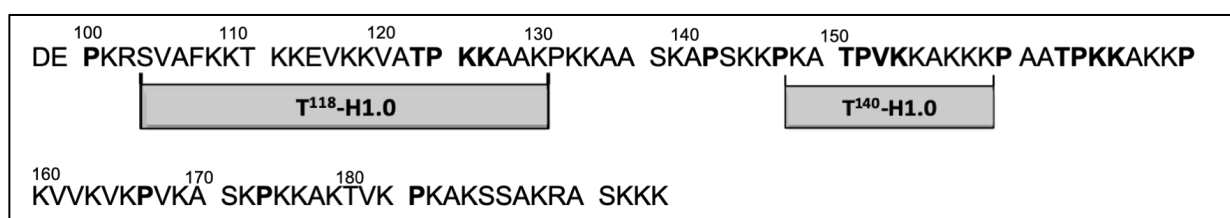
Figure 1. **Sequence of the C-terminal domain of histone H1.0 (C-H1.0).** Phosphorylation motifs are in bold and Pro residues underlined. Residues included in peptides T^{118}-H1.0 and T^{140}-H1.0 are indicated.

4.3.   RESULTS AND DISCUSSION

Peptide design

We designed two model peptides containing a single phosphorylation site of the three present in C-H1.0 (**Figure 1**) and a single Pro residue, to avoid multiple *cis*/*trans* isomerism equilibria that would complicate NMR analysis. Peptide T^{140}-H1.0 (Ac-KATPVKKAKKK-NH$_2$) was selected because it is the longest sequence containing the T^{140}PVK motif and no additional Pro residue (**Figure 1**), and peptide T^{118}-H1.0 (Ac-SVAFKKTKKEVKKVATPKKAAK-NH$_2$) as representative of the two TPKK phosphorylation motifs contained in C-H1.0 (T^{118}PKK and T^{152}PKK; **Figure 1**). Although CH-1 (**Figure 1**), a non-phosphorylated peptide spanning residues 99-121 and containing the T^{118}PKK motif, had been previously studied [11], we decided to design a new one (T^{118}-H1.0; **Figure 1**) in which the phosphorylation motif is not just the C-terminal end.

Thus, we included a few residues after the motif, as much as possible without including an additional Pro residue. On the other hand, some N-terminal residues present in peptide CH-1 were not included in peptide T^{118}-H1.0 to avoid having an additional Pro. These peptides and their phosphorylated counterparts, pT^{118}-H1.0 (Ac-SVAFKKTKKEVKKVA(pT)PKKAAK-NH$_2$) and pT^{140}-H1.0 (Ac-KA(pT)PVKKAKKK-NH$_2$), have their N- and C-termini, respectively, acetylated and amidated to avoid charged-end effects.

Effect of phosphorylation on Pro *cis/trans* isomerism

Since phosphorylation occurs at Pro-preceding Thr residues (T^{118}PKK and T^{140}PVK motifs; **Figure 1**), we examined whether the Pro *cis/trans* isomerism might be different in the non-phosphorylated and phosphorylated peptides (T^{118}-H1.0 *versus* pT^{118}-H1.0 and T^{140}-H1.0 *versus* pT^{140}-H1.0). Because of the Pro *cis/trans* isomerism, Pro-containing peptides typically contain two sets of NMR signals, one for the *trans* isomer and the other for the *cis*. Thus, peptide T^{118}-H1.0 in aqueous solution (**Figure S1**), and peptides T^{140}-H1.0 and pT^{140}-H1.0 in both aqueous solution and in 90 % TFE (**Figure S2**) show two sets of NMR signals. However, a single set of NMR signals was observed for peptide T^{118}-H1.0 in 90 % TFE and for pT^{118}-H1.0 in both aqueous solution and 90 % TFE. These signals as well as the major ones in the other cases correspond to the *trans* X-Pro isomer, as demonstrated by $\Delta\delta_{C\beta\gamma}$ values ($\Delta\delta_{C\beta\gamma}$ = $\delta_{C\beta} - \delta_{C\gamma}$, ppm) being in the range 4.0-4.6 ppm [12] (**Tables S1-S8**), and by the NOE between the Hα of the preceding Thr/pThr residue and the Pro H$_{\delta\delta'}$ protons. We fully assign ^{1}H and ^{13}C chemical shifts of the *trans*-species (Materials and Methods) of the four peptides in aqueous solution and in 90 % TFE (**Tables S1-S8**).

Concerning the *cis* species, residues K138 to V143 were assigned for peptides T^{140}-H1.0 and pT^{140}-H1.0 (**Tables S5-S8**). Based on the intensity ratios of equivalent *cis* and *trans* cross-peaks, the percentages of *cis* species were estimated to be 5 ± 2 % for T^{140}-H1.0 and 11 ± 4 % for pT^{140}-H1.0, both in aqueous solution at 5ºC, and 0.9 ± 0.2 % for T^{140}-H1.0 and 10 ± 5 % for pT^{140}-H1.0 in 90% TFE at 25 ºC (**Table S9**). For peptide T^{118}-H1.0 in aqueous solution, the only two minor observed signals were tentatively assigned to T118 and K120 (**Figure S1A**), and then the estimated percentage of *cis* is 3.2 ± 0.2 % (**Table S9**).

Thus, in the non-phosphorylated forms the percentage of *cis* X-Pro isomer is larger in peptide T[140]-H1.0 (T[140]PVK motif) than in peptide T[118]-H1.0 (T[118]PKK motif), which agrees with the percentage of *cis* being sequence-dependent. Upon phosphorylation, the peptide with the pT[140]PVK motif slightly increases the population of the *cis* X-Pro isomer in water and a little more in TFE, whereas become non-detectable in the case of the peptide with the pT[118]PKK motif. These slight differences in the *cis/trans* percentages might be biologically important by leading to differential interacting ways with DNA or any other biologically relevant partner.

Since the conformational equilibrium is strongly shifted toward the *trans* X-Pro isomers, we will refer exclusively to this from here on*.*

Structural behavior in aqueous solution

Once assigned the NMR spectra of the non-phosphorylated and phosphorylated peptides (**Tables S1-S8**), we examined the ^{1}Hα and ^{13}Cα conformational shifts ($\Delta\delta_{H\alpha} = \delta_{H\alpha}^{observed} - \delta_{H\alpha}^{RC}$, ppm; and $\Delta\delta_{C\alpha} = \delta_{C\alpha}^{observed} - \delta_{C\alpha}^{RC}$, ppm) in aqueous solution. The $\Delta\delta_{H\alpha}$ and $\Delta\delta_{C\alpha}$ values shown by peptides T[118]-H1.0 and pT[118]-H1.0 (**Figures 2A-B**) are almost identical and lie mostly within the random coil range ($|\Delta\delta_{H\alpha}| \leq 0.05$ ppm and $|\Delta\delta_{C\alpha}| \leq 0.4$ ppm). And the same similarity is observed if we compare the profiles of peptides T[140]-H1.0 and pT[140]-H1.0 (**Figure 3A-B**), whose $\Delta\delta_{H\alpha}$ and $\Delta\delta_{C\alpha}$ values are also mainly within the random coil range. The only residues displaying $\Delta\delta_{H\alpha}$ and $\Delta\delta_{C\alpha}$ values large in magnitude are T118 and pT118 in the T[118]-H1.0/pT[118]-H1.0 pair, and T140 and pT140 in the T[140]-H1.0/ pT[140]-H1.0 pair, which are those characteristics of Pro-preceding residues (positive $\Delta\delta_{H\alpha}$ in the range 0.15-0.30 ppm, and negative $\Delta\delta_{C\alpha}$ ranging from –2.5 to –1.7 ppm; [13].

In the two pairs of peptides the main, though very small, differences are located around the phosphorylated Thr (residue 118 in the pair T[118]-H1.0 pT[118]-H1.0, and 140 in the pair T[140]-H1.0 and pT[140]-H1.0), so that they might be ascribed to the effect of the pThr residue on the chemical shifts of their neighbours. The absence of non-sequential NOEs provides further confirmation of the peptides being mainly random coil in aqueous solution.

The fact that the CD spectra of peptides T[118]-H1.0, pT[118]-H1.0, T[140]-H1.0 and pT[140]-H1.0 in aqueous solution are mostly featureless except for a typical random coil strong minimum at about 195 nm (**Figure 4**) also agrees with the peptides being disordered in water.
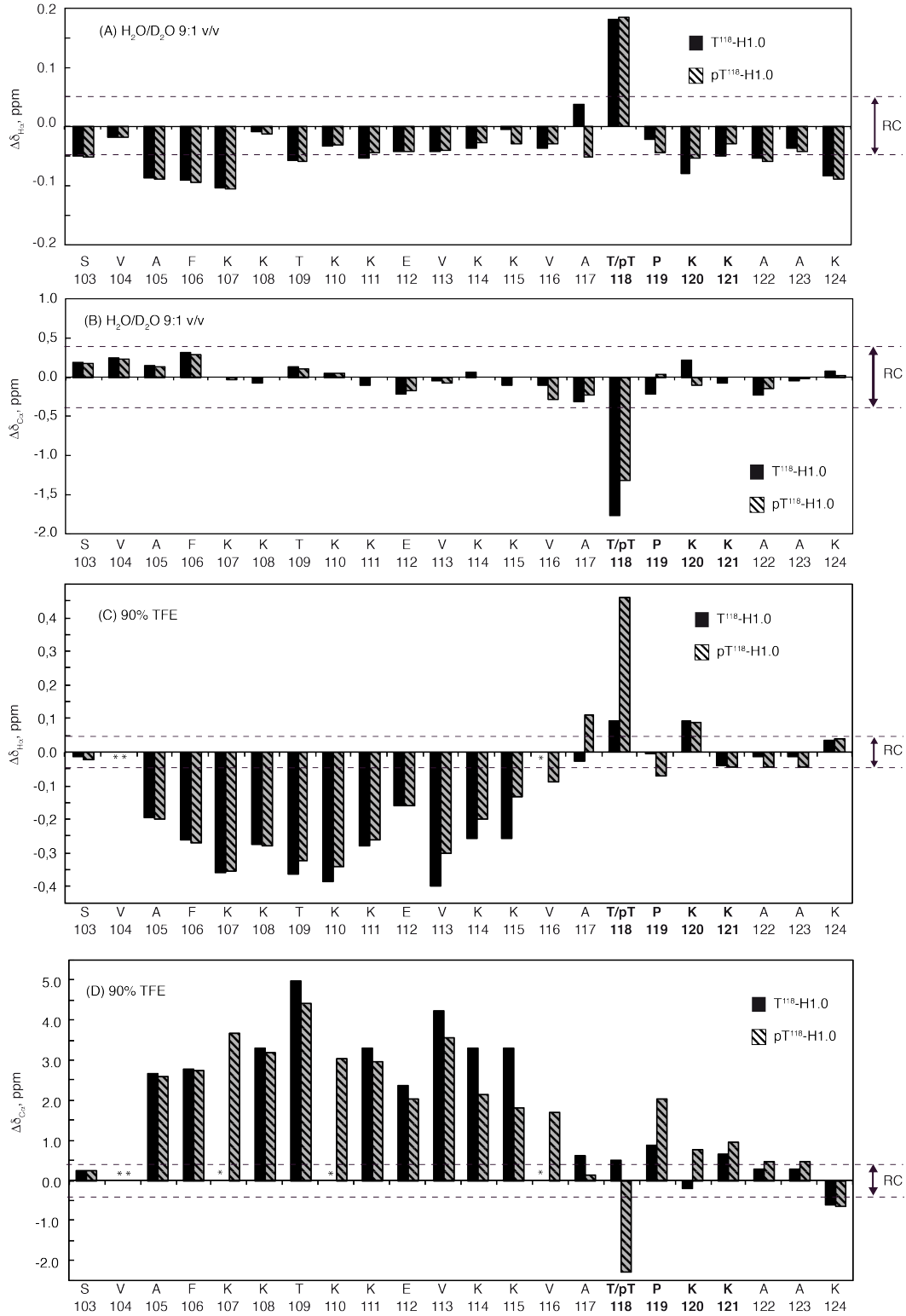
Figure 2. Conformational shifts for peptides T[118]-H1.0 (black bars) and pT[118]-H1.0 (grey stripped bars) as a function of sequence: (A) $\Delta\delta_{H\alpha}$ and (B) $\Delta\delta_{C\alpha}$ in aqueous solution at 5ºC, and (C) $\Delta\delta_{H\alpha}$ and (D) $\Delta\delta_{C\alpha}$ in 90 % TFE at 25ºC. In all panels, pH 5 and the two dashed lines indicate the random coil range (RC). An * indicates a non-observed value. Residues belonging to the phosphorylation motif are in bold.

This result is consistent with data from a non-phosphorylated peptide, which encompasses residues 99-121 of C-H1.0 (**Figure 1**) and contains the T[118]PKK phosphorylation site [11], as well as from the full-length C-H1.0 [3,10], which were shown to be disordered in aqueous solution. To get a more detailed comparison between the conformational behaviour of the sequences in the isolated peptides and within the full-length domain, we compare the chemical shifts of $^{13}C_\alpha$ and $^{13}C_\beta$ in the non-phosphorylated peptides with those in the non-phosphorylated C-H1.0 [10], and those in the phosphorylated peptides with those in the tri-phosphorylated pT-C-H1.0 [10]. The averaged differences $\Delta\delta$ ($\Delta\delta = \delta^{peptide} - \delta^{full\text{-}length\ domain}$, ppm) in $^{13}C_\alpha$ and $^{13}C_\beta$ are quite small (between –0.1 and +0.1 ppm; see Table S10), within the range of experimental errors. This result further confirms that the isolated peptides are able to reproduce the conformational behaviour they have within the full-length domain.
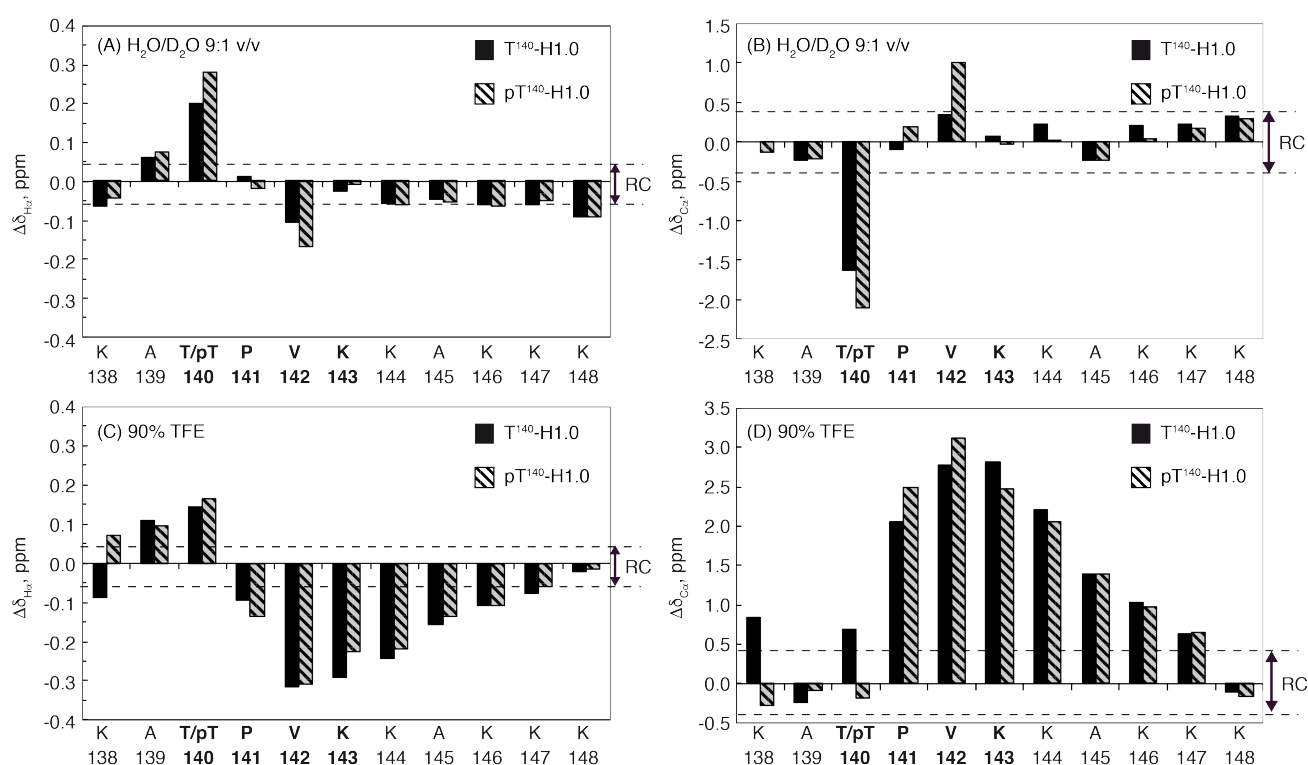
Figure 3. Conformational shifts for peptides T[140]-H1.0 (black bars) and pT[140]-H1.0 (grey stripped bars) as a function of peptide sequence: (A) $\Delta\delta_{H\alpha}$ and (B) $\Delta\delta_{C\alpha}$ for in aqueous solution, and (C) $\Delta\delta_{H\alpha}$ and (D) $\Delta\delta_{C\alpha}$ in 90 % TFE. In all panels, pH 5.5 and 5ºC, and the two dashed lines indicate the random coil range (RC). Residues belonging to the phosphorylation motif are in bold.
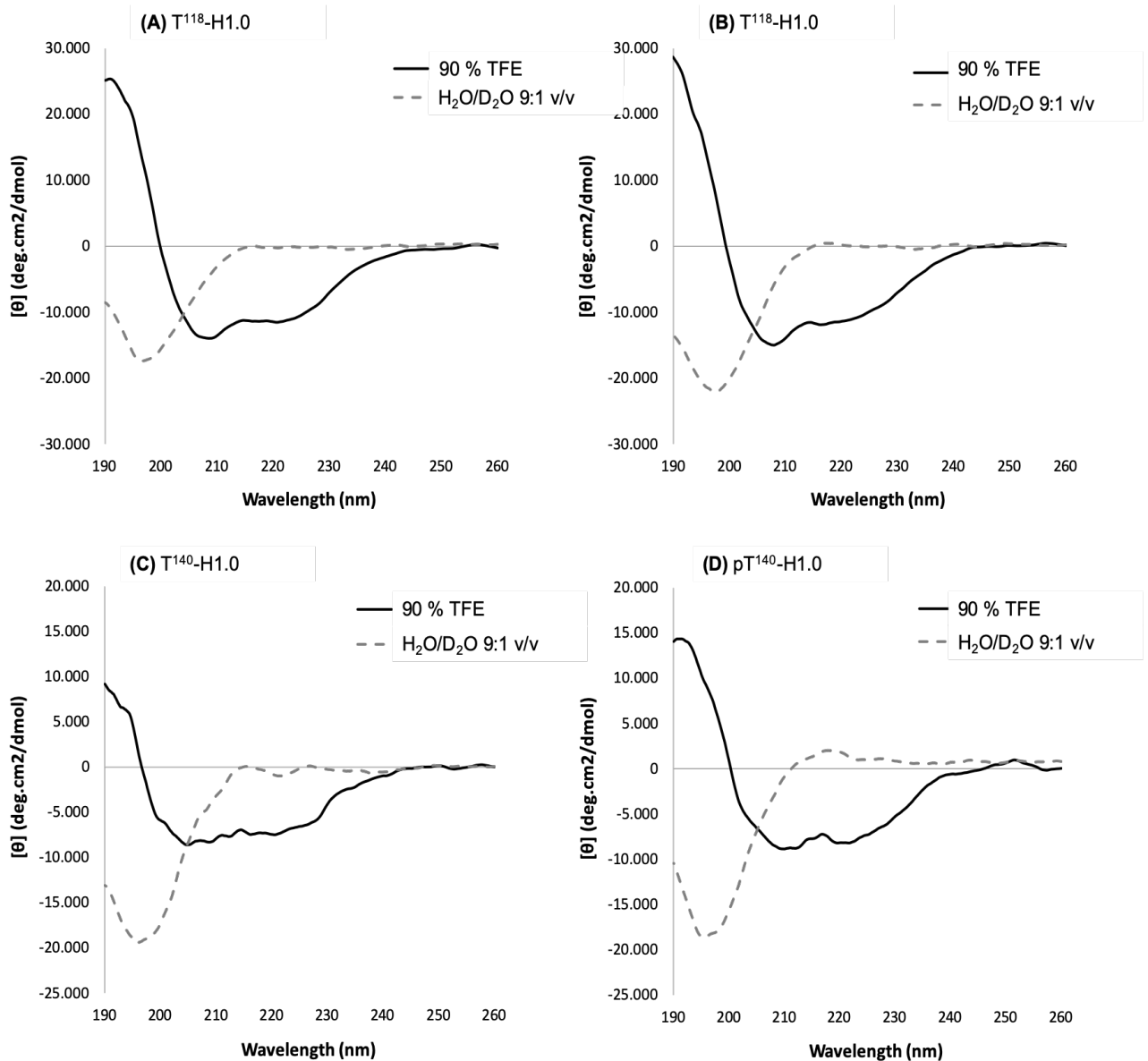
Figure 4. CD spectra of peptides T[118]-H1.0 (A), pT[118]-H1.0 (B), T[140]-H1.0 (C) and pT[140]-H1.0 (D) in aqueous solution (dotted line) and in 90% TFE (black line) at pH 5.5 and 25 ºC.

Table 1. Averaged $\Delta\delta$ values and α-helix populations estimated from $\Delta\delta_{H\alpha}$ and $\Delta\delta_{C\alpha}$ (see Methods) for peptides T[118]-H1.0, pT[118]-H1.0, T[140]-H1.0 and pT[140]-H1.0 in aqueous solution at pH 5.5 and 5ºC and in 90 %

TFE at pH 5.5 and 25 ºC. N- and C-terminal residues were excluded to calculate the averaged Δδ values. [a] Reported errors are standard deviations for the mean of the percentages obtained from $\Delta\delta_{H\alpha}$ and $\Delta\delta_{C\alpha}$ values. [b] Values at 25 ºC. [c] Values in brackets are at 5 ºC.

Peptide	Helix length	Conditions	$\Delta\delta_{H\alpha}$, ppm	% α-helix from $\Delta\delta_{H\alpha}$	$\Delta\delta_{C\alpha}$, ppm	% α-helix from $\Delta\delta_{C\alpha}$	Averaged % α-helix[a]
T[118]-H1.0	105-115	H_2O	-0.05	13	+0.12	4	9 ± 5
		90 % TFE[b]	-0.29	75	+3.35	100	87 ± 13
pT[118]-H1.0	105-115	H_2O	-0.05	13	+0.15	5	9 ± 4
		90 % TFE[b]	-0.27	66	+2.92	95	81 ± 14
T[140]-H1.0	141-147	H_2O	-0.06	16	+0.21	7	12 ± 5
		90 % TFE[b,c]	-0.15 (-0.18)	39 (47)	+1.58 (+1.92)	51 (62)	45 ± 6 (55 ± 6)
pT[140]-H1.0	141-147	H_2O	-0.06	16	+0.28	9	13 ± 4
		90 % TFE[b,c]	-0.16 (-0.16)[c]	40 (40)	+1.80 (+1.80)	58 (58)	49 ± 9 (49 ± 9)
C-H1.0	105-115	H_2O	---	---	+0.18	6	---
	141-147	H_2O	---	---	+0.25	8	---
pT-C-H1.0	105-115	H_2O	---	---	+0.12	4	---
	141-147	H_2O	---	---	+0.25	8	---

Structural behavior in the presence of trifluoroethanol (TFE)

According to CD data, C-H1.0 was shown to increase its helical content [2], and a previously studied non-phosphorylated peptide CH-1, which contains the T[118]PKK motif (**Figure 1**), becomes helical in the presence of TFE [11]. Therefore, we decide to examine if other peptides derived from the C-H1.0 would display similar behaviour, and more interestingly to see if the non-phosphorylated peptides (T[118]-H1.0 and T[140]-H1.0) and their phosphorylated counterparts (pT[118]-H1.0 and pT[140]-H1.0) behave similarly or differently in the presence of TFE.

To this end, we first examined the effect of TFE on the CD spectra of the four peptides. Upon TFE titration we observed a progressive conversion of the minima at 195 nm into a maximum at about 197 nm, indicating that the peptides acquire some ordered conformation. At 90 % TFE (**Figure 4**), the peptides show minima at positions close to those characteristic of helices (208 nm and 222 nm). To better characterise the TFE-stabilised structures, we proceeded to study the four peptides in 90 % TFE by NMR.

As in aqueous solution, the profiles of $\Delta\delta_{H\alpha}$ and $\Delta\delta_{C\alpha}$ values of the corresponding non-phosphorylated and phosphorylated peptides are practically identical (**Figures 2C-D and 3C-D**). But, in contrast to aqueous solution, the magnitudes of many $\Delta\delta_{H\alpha}$ and $\Delta\delta_{C\alpha}$ values are outside the random coil range. Thus, the pair of peptides T[118]-H1.0 and pT[118]-H1.0 shows a stretch of negative $\Delta\delta_{H\alpha}$ and positive $\Delta\delta_{C\alpha}$ values, large in magnitude, extending residues A105 to V116 (**Figure 2C-D**), which indicates that they form helical structures in that region. Analogously, the stretch of negative $\Delta\delta_{H\alpha}$ and positive $\Delta\delta_{C\alpha}$ values observed in T[140]-H1.0 and pT[140]-H1.0 shows that they form a helix spanning from P141 to K147. It is worth to note that the profiles of the four peptides in aqueous solution follow the same pattern that in 90 % TFE (same signs), except for their very small magnitudes (see above). According to this, these peptides in aqueous solution are mostly random coil, but there might exist a low population of helical structures, which would span the same residues that in the presence of TFE. Further and stronger evidence about the four peptides forming helical structures in 90 % TFE comes from the fact that they show some helix-characteristic NOEs, such as medium-range $\alpha N(i, i+2)$, $\alpha N(i, i+3)$, $\alpha N(i, i+4)$ and $\alpha\beta(i, i+3)$, and intense sequential $NN(i, i+1)$.

The helix populations can be estimated from the $\Delta\delta_{H\alpha}$ and $\Delta\delta_{C\alpha}$ averaged for the helical residues (**see Materials and Methods**). As seen in Table 1, the helix spanning residues 105-116 is highly populated in T[118]-H1.0 and pT[118]-H1.0 in 90% TFE (81-87 % at pH 5.5 and 25 ºC). The population of the helix formed by T[140]-H1.0 and pT[140]-H1.0 in 90% TFE is not so high (45-49 % at pH 5.5 and 25 ºC; **Table 1**), but increases at low temperature (53-55 % at 5 ºC; **Table 1**). Applying the same procedure the helix percentages present in the conformational ensemble equilibrium of these peptides in aqueous solution are very small, in the range 7-13 % at pH 5.5 and 5 ºC (**Table 1**). Based on the $\Delta\delta_{C\alpha}$ averages (obtained from the chemical shifts reported in Chapter 2 [10]) the helix percentages for the segments 105-115 and 141-147 in the full-length C-H1.0 and pT-C-H1.0 in aqueous solution are quite similar (4-

8 % at pH 5.5 and 25 ºC) to those in the isolated peptides. Comparing once again non-phosphorylated and phosphorylated forms, it is noticeable that the differences in helix populations between the two forms in the pairs T[118]-H1.0/pT[118]-H1.0, and T[140]-H1.0/pT[140]-H1.0 are in the range 1-6 %, which is within the experimental error (3-7 %) [14]. Overall this data is consistent with the peptides, and also the full-length C-terminal domain, existing in aqueous solution as disordered ensembles, which contain residual helical populations in segments 105-115 and 141-147. Such helices are greatly stabilised by TFE.

The features of the helices formed by the four peptides can be visualised by performing structure calculations on the basis of distance and angle restraints derived from the NMR parameters observed in 90% TFE (**see Methods and Table S11**). Structure calculations in aqueous solution would be meaningless because of the low helical populations (less than 15 %; **Table 1**) and the absence of any non-sequential NOE. The quality of the resulting structures in the four peptides is shown by their ϕ,ψ angles being in the allowed regions of the Ramachandran diagram (**Table S11**).

The resulting structure of T[118]-H1.0 (**Figure 5A-C and S3A**) is well defined (RMSD for the backbone atoms in the ensemble of the 20 lowest target function conformers is 0.7 ± 0.3 Å; **Table S9**), and shows two regular helices: a long α-helix extending residues V104-K115, approximately coincident with that identified by qualitative analysis of $\Delta\delta_{H\alpha}$ and $\Delta\delta_{C\alpha}$ values (see above), and a very short helix spanning approximately residues P119-A123, which is classified as α or 3_{10} depending on the conformer. These two helical regions are connected by a turn/loop at approximately residues A117-T118. The population of this second short helix, which was undetected by our qualitative analysis of $\Delta\delta_{H\alpha}$ and $\Delta\delta_{C\alpha}$ values because of its small magnitudes, which are mostly within the random coil range even in 90% TFE (**Figure 2**), would be of only 9 ± 4 % in 90% TFE.

The angle between the helices is relatively well defined, in the range 98º ± 8º. The long α-helix 104-115 (**Figure 5B**) exhibits an amphipathic character, having all Lys side chains pointing towards the same helix face, except for K108, which forms a salt bridge with the side chain of E112. This helix might be stabilised by the K108/E112 salt bridge and an interaction between the aromatic ring of F106 and the side chain of K110. Apart from their closeness in the calculated structure, the F106/K110 interaction is confirmed by the deviation from random coil values of some K110 side chain protons, which is due to anisotropy effects from the Phe aromatic ring.

This helical conformation might correspond to the residual helix existing in aqueous solution, and even be the one within the disordered ensemble of the C-terminal domain of histone H1.0, which is active for DNA interaction. Concerning the short C-terminal helix, it involves the phosphorylation motif (T[118]PKK), and T118 or/and P119 could be acting as N-caps. The conformation of this motif, if present in the disordered ensemble of the C-terminal domain, might facilitate access of the cyclin-dependent kinases to the Thr hydroxyl to be phosphorylated.

Like in the non-phosphorylated peptide T[118]-H1.0, the structure of the phosphorylated pT[118]-H1.0 shows two helices (**Figures 5D-F**), a long α-helix at segment 104-115 and a short α/3_{10} at residues 119-123, but their relative disposition is not defined. In fact, the structure of T[118]-H1.0 is not well defined when considering all its residues (**see Figure S3B**), being the RMSD for all backbone atoms (residues 104-123) quite large, 1.9 ± 1.0 Å. However, both helices by themselves are well defined, as shown by the RMSD for backbone atoms, which is 0.6 ± 0.2 Å for residues 104-115, and 0.3 ± 0.1 Å for residues 120-123 (**Table S9**). The long N-terminal helix in pT[118]-H1.0 is quite similar to that in the non-phosphorylated peptide, and the same stabilising interactions are observed (K108/E112 and F106/K110; Figure 5D-E). Figure 5F shows the good definition of the C-terminal short helix by itself, which also is quite similar to that in the non-phosphorylated peptide (**Figure 5C**). In this short helix, there exists an interaction between the phosphate group of pT118 and the side chain of K120.

Thus, the main difference between the structures of the non-phosphorylated and phosphorylated peptides lies in the relative orientation of the two helices, which is relatively fixed in T[118]-H1.0, and poorly defined in pT[118]-H1.0, as can be appreciated in **Figure 5G**, where the backbone atoms of the N-terminal α-helix (residues 104-115) of the two peptides are overlaid. There, it is clearly seen that the orientation of the short C-terminal helix (120-123) is different for each conformer. Thus, the angle between the two helices is in a wide range 45º ± 17º. Apart from being poorly defined, the relative disposition between the two helices is different in T[118]-H1.0 and pT[118]-H1.0 (**Figure 5G**). This conformational difference, together with the negative charge of the phosphate group, might be responsible for the lower affinity of the full-length C-H1.0 to bind DNA.
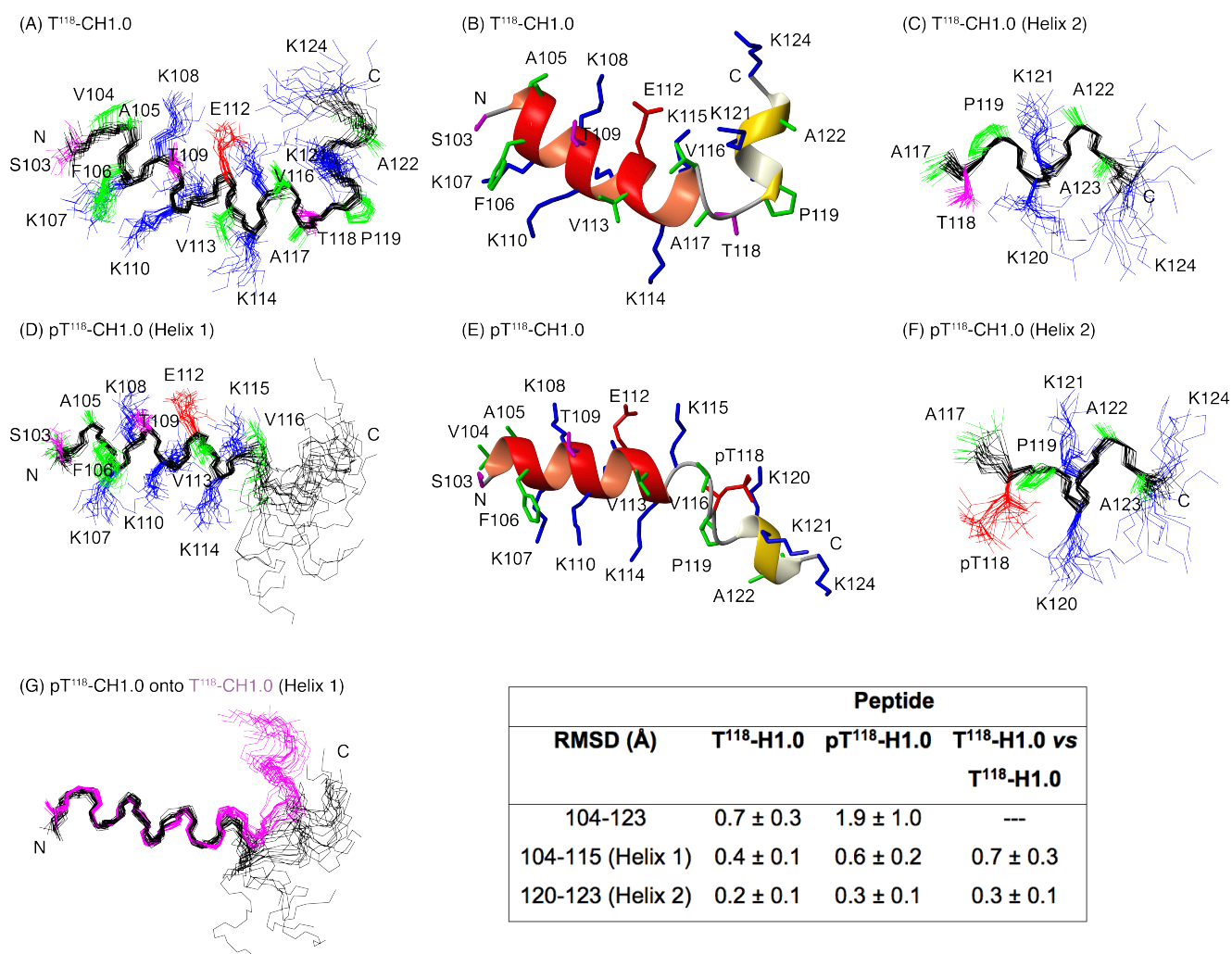
Figure 5. Structures of T[118]-H1.0 and pT[118]-H1.0 in 90% TFE. Structural ensemble of the 20 lowest target function conformers calculated for T[118]-H1.0 **(A)** and pT[118]-H1.0 **(B).** pT[118]-H1.0 superposition is done for backbone atoms of residues 104-115. Backbone atoms are in black, and side chains in blue for K, in green for non-polar residues (A, P and V), in magenta for S and T, and in red for pT. Representative structures for T[118]-H1.0 **(C)** and pT[118]-H1.0 **(D)** with the backbone shown as a ribbon and the side chains in neon. The N-terminal long helix is in red and the short C-terminal in gold. Superposition of the 20 lowest target function conformers for the backbone atoms of residues 119-124 for T[118]-H1.0 **(E)** and pT[118]-H1.0 **(F).** (G) Superposition of the backbone atoms for helix 1 in T[118]-H1.0 (in magenta) and pT[118]-H1.0 (in black). N- and C-termini are labelled in all the panels, and side chains in panels A-B and D-E. The inset gives the pairwise RMSD values for the backbone atoms of the structural ensembles.

(A) T^{140}-CH1.0

(B) pT^{140}-CH1.0

(C) T^{140}-CH1.0

(D) pT^{140}-CH1.0

(E) T^{140}-CH1.0 onto pT^{140}-CH1.0

Peptide	RMSD (Å)
T^{140}-H1.0	0.4 ± 0.1
pT^{140}-H1.0	0.5 ± 0.2
T^{140}-H1.0 *vs* pT^{140}-H1.0	0.6 ± 0.2

Figure 6. Structures of peptides T^{140}-H1.0 and pT^{140}-H1.0 in 90% TFE. **(A-B)** Structural ensemble of the 20 lowest target function conformers calculated for peptides T^{140}-H1.0 **(A)** and pT^{140}-H1.0 **(B).** Backbone atoms are in black, and side chains in blue for K, in green for non-polar residues (A, P and V), in magenta for T and in red for pT. **(C)** Superposition of the backbone atoms of the structural ensembles for T^{140}-H1.0 (in magenta) and pT^{140}-H1.0 (in black). **(D-E)** Representative structures for T^{140}-H1.0 **(D)** and pT^{140}-H1.0 **(E)** with the backbone shown as a ribbon and the side chains in neon. N- and C-termini are labelled in all the panels, and side chains in panels A-B and D-E. The inset gives the pairwise RMSD values for the backbone atoms of residues 139-147 of the structural ensembles.

Concerning the structures of T^{140}-H1.0 and pT^{140}-H1.0, they showed well-defined α-helices spanning residues 141-147 (**Figure 6A-D; Table S9**). Considering the 40 con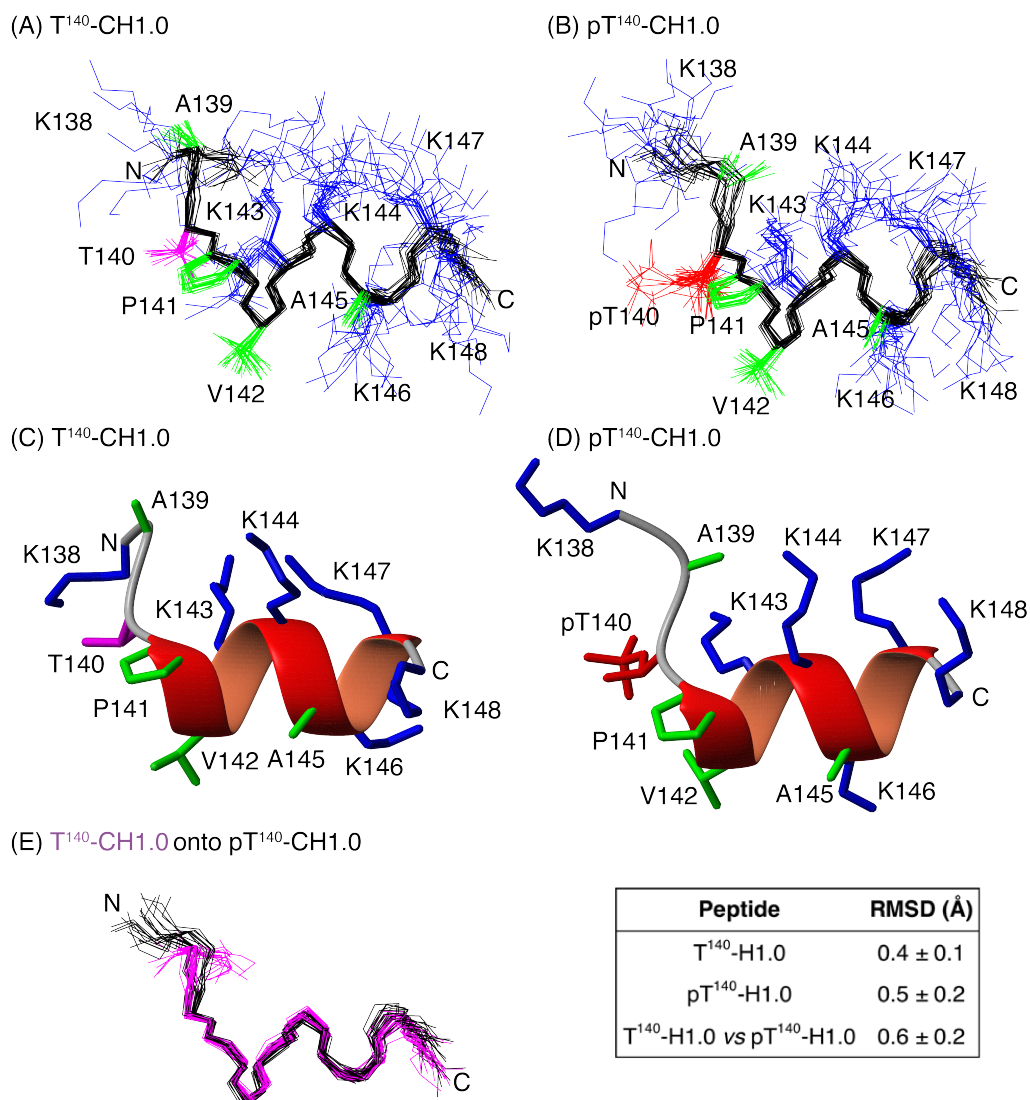formers, from both ensembles, the RMSD value raises only slightly to 0.6 ± 0.2 Å, demonstrating that, in this case, the structures adopted by the non-phosphorylated and the phosphorylated peptides are very similar (**Figure 6E**). The side chain of residue pT140 seems to interact with the side chain of K143 (**Figure 6D**).

It is interesting to compare the structure of the phosphorylation T^{118}PKK motif present in peptides T^{118}-H1.0/pT^{118}-H1.0 with that of the T^{140}PVK motif present in peptides T^{140}-H1.0/pT^{140}-H1.0. The backbone structure is pretty similar in the four peptides, likely determined by the Pro residue. However, Thr and pThr residues show different interactions with the i+2 and i+3 residues. Thus, in the structures of peptides T^{118}-H1.0 and pT^{118}-H1.0, T118 and pT118 interacts with the side chain of the i+2 Lys, but not with the i+3 Lys (see T118/pT118 and K120 in Figure 5F), but T140 and pT140 in peptides T^{140}-H1.0 and pT^{140}-H1.0 interacts with the side chain of the i+3 Lys, more than with the i+2 Val (see pT140 and K143 in Figure 6D). These differences might be biologically relevant and be related to different roles for the various phosphorylation motifs present in the C-terminal domain of Histone H1.0.

4.4. CONCLUSIONS

We have examined the structural behaviours of two C-H1.0-derived peptides, T^{118}-H1.0 and T^{140}-H1.0, which contain phosphorylation motifs TPKK and TPVK, respectively, and compared them with those in their phosphorylated counterparts (pT^{118}-H1.0 and pT^{140}-H1.0). Given that the phosphorylation motifs contain a Pro residue, we analysed the *cis/trans* isomerism finding that the *cis* percentages are different in the T^{118}PKK and T^{140}PVK motifs, and that they diminishes in the T^{118}PKK motif and increments in the T^{140}PVK motif. In any case, the *cis* percentages are quite low (< 10%), so that our NMR study was focussed on the structural behaviour of the major *trans* species. It should be mentioned that no information about the *cis/trans* isomerism could be obtained from the NMR study of the full-length C-H1.0, whose sequence contains 12 Pro residues in total (Chapter 2; [10]).

CD and NMR data evidenced that the four peptides are mainly random coil in aqueous solution, but they become structured in the presence of 90 % TFE. The fact that phosphorylation does not affect peptide structural behaviour in aqueous solution is in agreement with previous results on the full-length C-terminal H1.0 domain (C-H1.0), which is intrinsically disordered and mainly unaffected by phosphorylation [10]. Hence, this result validates the usefulness of minimalist approaches based on model peptides to obtain structural info about intrinsically disordered proteins, particularly if they are short-lived, and so difficult to study by NMR.

In the presence of TFE, the structures of T^{118}-H1.0 and its phosphorylated counterpart pT^{118}-H1.0 consist of two helical regions: a long amphipathic α-helix (residues 104-115) and a short $α/3_{10}$ helix (residues 119-123). But, the orientation between the helices is well defined, approximately perpendicular, in the non-phosphorylated T^{118}-H1.0, and poorly defined, and not perpendicular, in the phosphorylated pT^{118}-H1.0. The structures formed by peptides T^{140}-H1.0 and pT^{140}-H1.0 are very similar α-helices (residues 141-147). That the structures formed by these peptides are helical agrees with CD data on the full-length C-H1.0, which indicated helix formation in the presence of TFE [3].

Focussing in the structural features of the phosphorylation motifs, the TPKK and TPVK sequences displayed the same backbone conformation in the two peptides and in both the non-phosphorylated and phosphorylated states. Interestingly, side-chain contacts are different. Thus, Thr and pThr side chains interact with the i+2 Lys side chain in the structures of peptides T^{118}-H1.0 and pT^{118}-H1.0 (TPKK motif), and with the i+3 Lys side chain in the case of peptides T^{140}-H1.0 and pT^{140}-H1.0 (TPVK motif). These differences might be biologically relevant and be related to different roles for the various phosphorylation motifs present in the C-terminal domain of Histone H1.0.

Finally, it is important to highlight that short model peptides (11 and 22 residues) are found able to reproduce the conformational behaviour of the full-length C-H1.0 domain. Furthermore, the NMR study of the model peptides have provide us structural details, which were very difficult to get by studying the full-length C-H1.0 domain, because of its very repetitive sequence and its short-life. In brief, minimalist approaches are effective as alternative and/or complementary strategies to get experimental data on IDPs.

4.5. MATERIALS AND METHODS

Materials and peptide synthesis

The deuterated compounds [D2]-TFE (99.0 %), and D_2O (99.9 %) were from Cambridge Isotope Lab (USA). Peptides were synthesised using Fmoc (fluorenyl-9-methyloxycarbonyl) solid phase protocols and purified by reverse-phase HPLC up to 95 % or more purity by Caslo Aps (Lyngby, Denmark).

T^{118}-H1.0 (Ac-SVAFKKTKKEVKKVATPKKAAK-NH$_2$): RP-HPLC: t_R = 13.89 min; 98.35 % (linear 5-25 % B gradient in 20 min; buffer A: 0.05 % TFA in H_2O/CH_3CN 98:2; buffer B: 0.05 % TFA in H_2O/CH_3CN 1:9). HRMS: Theoretical molecular weight (MW) = 2457.07; Found $[M+H]^+$ = 2457.84.

pT^{118}-H1.0 (Ac-SVAFKKTKKEVKKVA(pT)KKAAK-NH$_2$): RP-HPLC: t_R = 10.29 min; 99.38 % (linear 8-25 % B gradient in 17 min; buffer A: 0.05 % TFA in H_2O/CH_3CN 98:2; buffer B: 0.05 % TFA in H_2O/CH_3CN 1:9). HRMS: Theoretical molecular weight (MW) = 2537.05; Found $[M+H]^+$ = 2538.13.

T^{140}-H1.0 (Ac-KATPVKKAKKK-NH$_2$): RP-HPLC: t_R = 11.46 min; 100 % (linear 0-17 % B gradient in 17 min; buffer A: 0.05 % TFA in H_2O/CH_3CN 98:2; buffer B: 0.05 % TFA in H_2O/CH_3CN 1:9). HRMS: Theoretical molecular weight (MW) = 1267.64; Found $[M+H]^+$ = 1267.73.

pT^{140}-H1.0 (Ac-KA(pT)PVKKAKKK-NH$_2$): RP-HPLC: t_R = 9.70 min; 95.1 % (linear 0-15 % B gradient in 15 min; buffer A: 0.05 % TFA in H_2O/CH_3CN 98:2; buffer B: 0.05 % TFA in H_2O/CH_3CN 1:9). HRMS: Theoretical MW = 1347.62; Found $[M+H]^+$ = 1348.20.

Circular Dichroism

All experiments were recorded in a Jasco J-810 spectropolarimeter equipped with a Peltier temperature control unit. CD spectra of the peptides were carried out in aqueous solution and in different percentages of TFE at pH 5.5 at peptide concentrations of approximately 37 µM. Since the peptides lack aromatic residues, concentrations were estimated by weighting the peptide to prepare a 370 µM stock solution. Cell path lengths of 0.1 cm were used.

Ellipticity was recorded at 220 nm at a temperature of 25 ºC in all cases.

NMR experiments

Samples for NMR spectra acquisition were prepared at 0.5-1.0 mM concentration and contained sodium 2,2-dimethyl-2-silapentane-5-sulfonate (DSS) as internal reference for ^{1}H chemical shifts. In the case of peptides, the required amount of lyophilised peptide was solved in either H_2O/D_2O 9:1 v/v or in 90 % [D2]-TFE in H_2O/D_2O 9:1 v/v. In the case of DNA, the samples were prepared at 0.05 mM concentration. The peptide was solved in H_2O/D_2O 9:1 v/v and 10 mM NaCl at pH 5.5. Minimal amounts of NaOD or DCl were used to adjust pH, which was measured with a glass micro-electrode and not corrected for isotopic effects.

NMR spectra were recorded on a Bruker Avance-600 spectrometer operating at a proton frequency of 600.1 MHz and equipped with a cryoprobe. A methanol sample was employed to calibrate cryoprobe temperature. As formerly reported [15], 1D ^{1}H NMR and 2D phase-sensitive two-dimensional correlated spectroscopy (COSY), total correlated spectroscopy (TOCSY), nuclear Overhauser enhancement spectroscopy (NOESY), and ^{13}C natural abundance [1]H-^{13}C heteronuclear single quantum coherence (HSQC) and ^{15}N natural abundance ^{1}H-^{15}N-HSQC spectra were recorded by standard techniques at 5 and 25 ºC, and processed using the TOPSPIN program (Bruker Biospin, Karlsruhe, Germany). Water signal was suppressed by presaturation. [13]C and [15]N δ-values were indirectly referenced using the IUPAC-IUB recommended ^{1}H/[13]C (0.25144953) and ^{1}H/[15]N (0.101329118) chemical shift ratios [16].

Assignment of ^{1}H chemical shifts for the peptides was performed by analyses of the 2D NMR spectra using the SPARKY software (T. D. Goddard and D. G. Kneller, SPARKY 3, University of California, San Francisco) and following the standard sequential assignment strategy [17,18]. The ^{13}C chemical shifts were assigned from the ^{1}H/^{13}C cross-peaks present in the ^{1}H,^{13}C-HSQC spectra, and ^{15}N chemical shifts from the ^{1}H/^{15}N correlations observed in ^{1}H,^{13}C-HSQC spectra. Tables S1-S4 list the ^{1}H, ^{13}C and ^{15}N chemical shifts for peptides T[140]-H1.0 and pT[140]-H1.0 in aqueous solution and in 90 % TFE.

The [1]H_α and [13]C_α conformational shifts are defined by equations [1] and [2], respectively:

$$\Delta\delta_{H\alpha} = \delta_{H\alpha}^{observed} - \delta_{H\alpha}^{RC}, ppm \qquad \text{Eq. [1]}$$

$$\Delta\delta_{C\alpha} = \delta_{C\alpha}^{observed} - \delta_{C\alpha}^{RC}, ppm \qquad \text{Eq. [2]}$$

where the reference $\delta_{H\alpha}^{RC}$ and $\delta_{C\alpha}^{RC}$ values for the random coil (RC) state are taken from [13].

α-helix populations were estimated from ^{1}Hα and ^{13}Cα chemical shifts by a previously described method, which assumes a two-state α-helix /unfolded transition and applies equations [3] and [4] for ^{1}Hα and ^{13}Cα, respectively.

$$\% \, \alpha - helix = \frac{\langle \Delta\delta_{H\alpha} \rangle}{-0.39} \times 100 \qquad \text{Eq. [3]}$$

$$\% \, \alpha - helix = \frac{\langle \Delta\delta_{C\alpha} \rangle}{3.09} \times 100 \qquad \text{Eq. [4]}$$

being $\langle \Delta\delta_{H\alpha} \rangle$ and $\langle \Delta\delta_{C\alpha} \rangle$, respectively, the $\Delta\delta_{H\alpha}$ and $\Delta\delta_{C\alpha}$ values averaged for all the helical residues, as defined by equations [5] and [6]

$$\langle \Delta\delta_{H\alpha} \rangle = \sum_i \frac{\Delta\delta_{H\alpha}^i}{n} \qquad \text{Eq. [5]}$$

$$\langle \Delta\delta_{C\alpha} \rangle = \sum_i \frac{\Delta\delta_{C\alpha}^i}{n} \qquad \text{Eq. [6]}$$

where the $\Delta\delta_{H\alpha}^i$ and $\Delta\delta_{C\alpha}^i$ are, respectively, the $\Delta\delta_{H\alpha}$ and $\Delta\delta_{C\alpha}$ conformational shifts for residue i (residues 3 to 13 in peptides T[118]-H1.0 and pT[118]-H1.0, and residues 4 to 10 in peptides T[140]-H1.0 and pT[140]-H1.0), and n the number of helical residues (11 in peptides T[118]-H1.0 and pT[118]-H1.0, and 7 in peptides T[140]-H1.0 and pT[140]-H1.0).

Assuming experimental errors in the measurement of ^{1}H and ^{13}C δ-values of ± 0.01 ppm and ± 0.1 ppm, respectively, the errors in the estimated populations are ± 3 % and ± 7 %.

Structure calculation

Structures for peptides T[118]-H1.0 and pT[118]-H1.0, T[140]-H1.0 and pT[140]-H1.0 in 90 % TFE were calculated using the standard iterative protocol for automatic NOE assignment of the CYANA 2.1 program [19]. This procedure carries out seven cycles of combined automated NOE assignment and structure calculation, in which 100 conformers were calculated per cycle. The experimental input data were the lists of: (1) assigned chemical shifts, (2) NOE integrated cross-peaks present in 150 ms NOESY spectra, and (3) ɸ and ψ dihedral angle restraints. The automatic integration subroutine of SPARKY software (T. D. Goddard and D. G. Kneller, SPARKY 3, University of California, San Francisco) was used to integrate NOE cross-peaks.

Restraints for dihedral angles were obtained from ^{1}H, ^{13}C and ^{15}N chemical shifts using TALOSn webserver [20]. The non-standard phosphorylated Thr was incorporated into the CYANA amino acid library from that built by [21]. For each peptide, the final structure is the ensemble of the 20 lowest target function conformers calculated at the final cycle. These ensembles were visualized and examined using the program MOLMOL [22].

4.6. ACKNOWLEDGMENTS

4.7. REFERENCES

1. Roque A, Ponte I, Suau P. Interplay between histone H1 structure and function. Biochim Biophys Acta . 2015/09/30. 2016;1859(3):444–54. Available from: http://www.ncbi.nlm.nih.gov /pubmed/26415976

2. Vila R, Ponte I, Collado M, Arrondo JL, Jimenez MA, Rico M, et al. DNA-induced alpha-helical structure in the NH2-terminal domain of histone H1. J Biol Chem . 2001/10/05. 2001;276(49):46429–35. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11584004

3. Roque A, Iloro I, Ponte I, Arrondo JL, Suau P. DNA-induced secondary structure of the carboxyl-terminal domain of histone H1. J Biol Chem . 2005/07/12. 2005;280(37):32141–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16006555

4. Roque A, Ponte I, Suau P. Post-translational modifications of the intrinsically disordered terminal domains of histone H1: effects on secondary structure and chromatin dynamics. Chromosoma. 2016/04/22. 2017;126(1):83–91. Available from: http://www.ncbi.nlm.nih.gov /pubmed/27098855

5. Millán-Ariño L, Izquierdo-Bouldstridge A, Jordan A. Specificities and genomic distribution of somatic mammalian histone H1 subtypes. Biochim Biophys Acta - Gene Regul Mech . 2016;1859(3):510–9. Available from: http://dx.doi.org/10.1016/j.bbagrm.2015.10.013

6. Lu X, Hansen JC. Revisiting the structure and functions of the linker histone C-terminal tail domain. Biochem Cell Biol . 2003 Jun 1;81(3):173–6. Available from: https://doi.org/10.1139/o03-041

7. Roque A, Ponte I, Arrondo JL, Suau P. Phosphorylation of the carboxy-terminal domain of histone H1 : effects on secondary structure and DNA condensation. Nucleic Acids Res. 2008;36(14):4719–26.

8. Raghuram N, Strickfaden H, McDonald D, Williams K, Fang H, Mizzen C, et al. Pin1 promotes histone H1 dephosphorylation and stabilizes its binding to chromatin. J Cell Biol . 2013/10/09. 2013;203(1):57–71. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24100296

9. Lopez R, Sarg B, Lindner H, Bartolomé S, Ponte I, Suau P, et al. Linker histone partial phosphorylation: Effects on secondary structure and chromatin condensation. Nucleic Acids Res. 2015;43(9):4463–76.

10. Chaves-Arquero B, Pantoja-Uceda D, Roque A, Ponte I, Suau P, Jiménez MA. A CON-based NMR assignment strategy for pro-rich intrinsically disordered proteins with low signal dispersion: the C-terminal domain of histone H1.0 as a case study. J Biomol NMR. 2018;72(3–4):139–48. Available from: http://dx.doi.org/10.1007/s10858-018-0213-2

11. Vila R, Ponte I, Jimenez MA, Rico M, Suau P. A helix-turn motif in the C-terminal domain of histone H1. Protein Sci. 2000/05/04. 2000;9(4):627–36. Available from: http://www.ncbi.nlm.nih.gov/pubmed /10794405

12. Schubert M, Labudde D, Oschkinat H, Schmieder P. A software tool for the prediction of Xaa-Pro peptide bond conformations in proteins based on 13C chemical shift statistics. J Biomol Nmr. 2002/12/24. 2002;24(2):149–54. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12495031

13. Wishart DS, Bigam CG, Holm A, Hodges RS, Sykes BD. 1 H, 13 C and 15 N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects. J Biomol NMR. 1995;5(1):67–81.

14. Santiveri CM, Rico M, Jiménez MA. 13Cα and 13Cβ chemical shifts as a tool to delineate β-hairpin structures in peptides. J Biomol NMR. 2001;19(4):331–45.

15. Zamora-carreras H, Maestro B, Strandberg E, Ulrich AS, Sanz JM, Jimønez M. Micelle-Triggered b -Hairpin to a -Helix Transition in a 14-Residue Peptide from a Choline-Binding Repeat of the Pneumococcal Autolysin LytA. 2015;8076–89.

16. Markley JL, Bax A, Arata Y, Hilbers CW, Kaptein R, Sykes BD, et al. Recommendations for the presentation of NMR structures of proteins and nucleic acids IUPAC-IUBMB-IUPAB inter-union task group on the standardization of data bases of protein and nucleic acid structures determined by NMR spectroscopy. 5.

17. Wuthrich K. NMR of proteins and nucleic acids. The George Fisher Baker non-resident lectureship in chemistry at Cornell Unversity (USA). Wiley; 1986.

18. Kurt W, Martin B, Braun W. Polypeptide Secondary Structure Determination by Nuclear Magnetic Resonance Observation of Short Proton-Proton Distances. 1984;715–40.

19. Güntert P. Automated NMR Structure Calculation With CYANA. In: Downing AK, editor. Protein NMR Techniques. Totowa, NJ: Humana Press; 2004. p. 353–78. Available from: https://doi.org/10.1385/1-59259-809-9:353

20. Shen Y, Bax A. Protein backbone and sidechain torsion angles predicted fromNMR chemical shifts using artificial neural networks. J Biomol NMR. 2013;56(3):227–41.

21. Craft JW, Legge GB. An AMBER/DYANA/MOLMOL Phosphorylated Amino Acid Library Set and Incorporation into NMR Structure Calculations. J Biomol NMR. 2005 Sep;33(1):15–24. Available from: https://doi.org /10.1007/s10858-005-1199-0

22. Koradi R, Billeter M, Wiithrich K. MOLMOL : A program for display and analysis of macromolecular structures. 1996;7855(96):51–5.

CHAPTER 5

Structural characterization of the intrinsically disordered N-terminal domain of yeast eIF4G1 (Tif4631)

5.1. ABSTRACT

The intrinsically disordered N-terminal domain of eukaryotic translation initiation factor eIF4G1 $_{1-250}$ plays a key role in translation initiation by acting as a hub protein: it recognise various proteins and RNA . This chapter presents an extensive NMR study about the conformational properties of the free state of eIF4G1 $_{1-250}$. The NMR spectra of this large IDP has been assigned combining standard triple resonance methods with amino acid-selective unlabelling and isotope discrimination spectroscopy, an approach which greatly simplifies the spectra. The qualitative analysis of the chemical shifts, NOE data, ^{15}N relaxation, RDCs and three PRE dataset shows the presence of a highly populated α-helix in the BOX3 element and possible long-range contacts involving other conserved elements. To obtain a more quantitative description of the eIF4G1 $_{1-250}$ system a structural ensemble was calculated. A novel strategy was used combining experimental data with knowledge-based restraints that were built as ambiguous cation-π and π-π interactions. A very large pool of conformers (40000) was calculated using molecular dynamics and later ranked using a new ensemble building protocol guided by the experimental PRE data. The final ensemble is structurally very diverse and reproduce the experimental NMR data with very good agreement. The analysis shows the presence of non-specific cation-π and π-π networks that are nucleated around BOX3 α-helix. The eIF4G1 $_{1-250}$ ensemble represents one of the first examples of an IDP structure, and reproduces genuine characteristics of these type of molecules.

5.2. INTRODUCTION

Eukaryotes primary regulate translation at the level of initiation. Eukaryotic translation initiation factor eIF4G plays a key functional role in the cap-dependent translation initiation [1]. It works as the scaffold protein of eIF4F complex, which also includes eIF4E (cap-binding protein) and eIF4A (ATP-dependent RNA helicase) (**Figure 1A**) and as a hub protein for interaction with poly(A)-binding proteins, eIF3, etc. eIF4F is essential for the recruitment of 43S ribosome pre-initiation complex to mRNA template [2–4]. Moreover, eIF4G also controls the IRES-mediated translation [5]. In addition, several experiments indicate that eIF4G connects the nuclear cytoplasmic and life of mRNA. This notion received initial support from: (i) eIF4G shuttles between the cytoplasm and nucleus, and (ii) it interacts with several key nuclear factors such as the nuclear mRNA cap-binding complex or CBC. In the cytoplasm, CBC complex will be replaced by eIF4E (bound to the CAP structure of 5'-mRNA) and eIF4A (helicase that will allow the binding of mRNA to the ribosome) (**Figure 1A**). All these reasons support the notion that eIF4G is a crucial regulator of gene expression at multiple levels.

eIF4G also binds RNA [6,7] and RNA binding proteins like poly(A) and poly(U)-binding proteins (Pab1 and Pub1) [8,9]. These interactions makes eIF4G an essential component of stress granules structures [10–12]. Two isoforms of eIF4G proteins have been characterized from *Saccharomyces cerevisiae*, eIF4G1 (Tif4631) and eIF4G2 (Tif4632), which share 51 % sequence identity and 72 % similarity and are also related to human eIF4G counterparts (53 % identity and 46 % similarity) [13]. The eIF4G1 isoform is expressed at higher levels [6,14–16]. Biochemical studies revealed that eIF4G1 possesses several conserved domains, are also found in the mammalian isoforms (**Figure 1B**) [6]. The N-terminal of yeast eIF4G1 is a 400-residue low complexity domain (LCDs) predicted to be an IDD and with several conserved sequences (**Figure 1C**) that are thought to be molecular recognition features (MoRFs). These sequences typically function as recognition sites of proteins [14,17], RNAs [18] and also the recognition site for post-transcriptional modifications (PTMs). All these functions are important for the regulation of translation at several levels. More specifically the key conserved regions are concentrated within the first 250 amino acids: three boxes of about 15 – 20 residues and an RNA binding region (RNA1) (**Figure 1C**) [19].
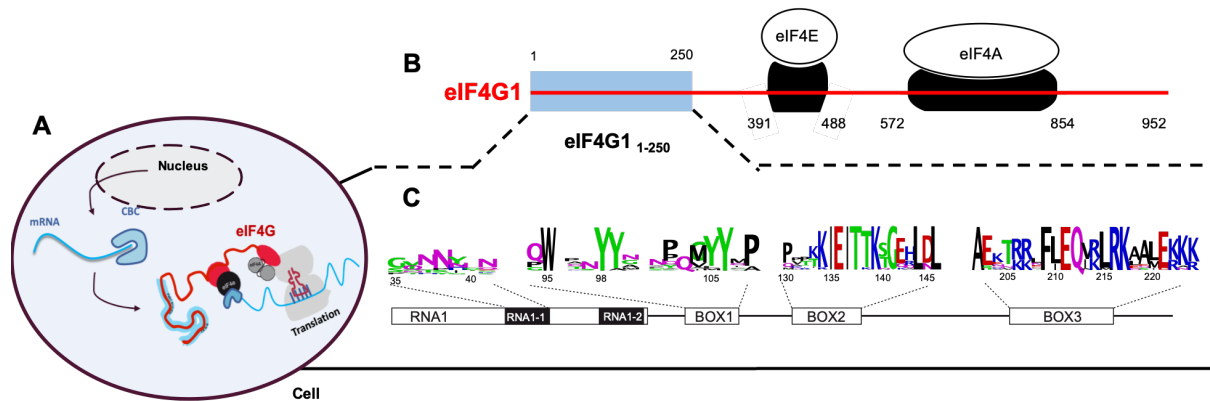
Figure 1. (A) Schematic representation of *S. Cerevisiae* eIF4G1 showing the interaction domains for the other components of the eIF4F trimer. **(B)** N-terminal domain (eIF4G1 $_{1-250}$) containing the conserved sequence boxes. The degree of conservation for each box across *S. Cerevisiae* is represented by weblogos.

The eIF4G1 $_{1-250}$ construct expressed at high levels, is easy to obtain in a pure form an, is stable for days under different pH conditions, in contrast to shorter forms like eIF4G1 $_{1-190}$ (**Figure S1**). NMR studies of large IDD like eIF4G1 $_{1-250}$ encounter difficulties due to high degree of signal overlap. C-detected methods tend to alleviate this, but their sensitivity is still lower than the proton-detected experiments. Other alternatives to simplify the spectra rely on sample production methods using strategies like amino acid-selective isotope labelling which simplifies the assignment of large IDPs sequences. The assignment of eIF4G1 $_{1-250}$ relies on previous assignments of shorter forms [20], new sets of standard triple resonance NMR experiments and isotope discrimination experiments [21,22] preformed on various samples with specific amino acid-selective isotope unlabelling in a ^{13}C/^{15}N background.

The structural studies of IDP/IDD are challenging due to: (1) the difficulty to obtain experimental restraints that represent individual conformations (most biophysical methods information averaged across all conformers in the sample). And (2) the lack of efficient computational protocols to generate structurally diverse ensembles. The traditional structural biology structure calculation methods use experimental restraints potentials to guide the molecular dynamics process. Using the same approach for the IDD implies that the experimental potential has to be evaluated across the whole ensemble, which might include at least thousands of individual structures, and for each dynamics step; increasing the computational cost enormously. In an effort to shortcut these two problems we have used a new strategy for the structural calculation of eIF4G1 $_{1-250}$ protein.

A large number of conformers were calculated by restrained molecular dynamics, combining experimental restrains (chemical shifts, 3D NOEs, ^{15}N relaxation data, residual dipolar couplings (RDC) and various set of paramagnetic relaxation enhancements (PRE)) with ambiguous cation-π and π-π interactions that are proposed to be crucial in IDDs [23] (knowledge-based restrains). The final eIF4G1 $_{1\text{-}250}$ NMR structure represents one of the first molecular descriptions of an IDD. The results expand our structural understanding of IDPs and open the way to new methodologies to calculate and analyse their structures.

5.3.   RESULTS AND DISCUSSION

eIF4G1 $_{1\text{-}250}$ assignment was achieved by amino acid-selective isotope labeling strategy

NMR is especially well-suited to study large IDP/IDD such eIF4G1$_{1\text{-}250}$ due to the favorable relaxation properties of these molecules. However, the spectral assignment of this protein was challenging due to the low signal dispersion and high overlap of its 2D ^{1}H-^{15}N HSQC spectrum. The complete assignment through standard triple resonance NMR experiments: 3D HNCO, 3D HNCA, 3D HC(CO)CA [24], 3D HN(CA)CO [25], 3D CBCANH [26] and 3D CBCA(CO)NH [27] was difficult. On the other hand, alternative assignment strategies like ^{13}C-detected methods, as shown in Chapter 2 of this thesis, proved to be not as sensible in this case. Particularly for regions having residual secondary structures and/or chemical exchange processes, that lower T_2. Alternatively, amino acid-selective isotope labeling of proteins is a powerful approach in protein NMR [28] because ^{15}N/^{13}C enrichment at selected residue types greatly reduce the number of signals in the spectra [29]. *Escherichia coli* has been the most commonly used host for heterologous production of foreign proteins [30], but selectively labeling cannot be simply achieved by adding individual ^{15}N and ^{13}C amino acid to the growth media, because metabolic dilution and scrambling of the isotopes [31]. These problems can be minimized by: (i) using *E. coli auxotrophic strains* with key enzymes of the amino acid biosynthetic pathways mutated /deleted [32], or (ii) by feeding these or standard strains with some certain amino acids precursors. In this sense, *E. coli auxotrophic strains* RF6 and RF10 has been used for production of Pro and Lys reverse-labelled eIF4G1 $_{1\text{-}250}$ samples.

The resulting recombinant samples contains Pro/Lys natural abundance residues in a $^{15}N/^{13}C$ uniformly labeled background. The standard 2D ^{1}H-^{15}N HSQC, shows the lacks of Lys peaks (in the reverse-Lys sample) and is identical for the reverse Pro (no HNs). However the real power of the technique becomes clear upon using isotope-discriminated 2D ^{1}H-^{15}N spectroscopy [22] that greatly simplifies the complexity of the spectra by only showing the peaks following (i) Lys or Pro residues (**Figure 2A-B**) in case of *E. coli auxotrophic*, and (ii) Ile, Leu and Val (**Figure 2C**) when using α-ketoacid precursors of these amino acids.

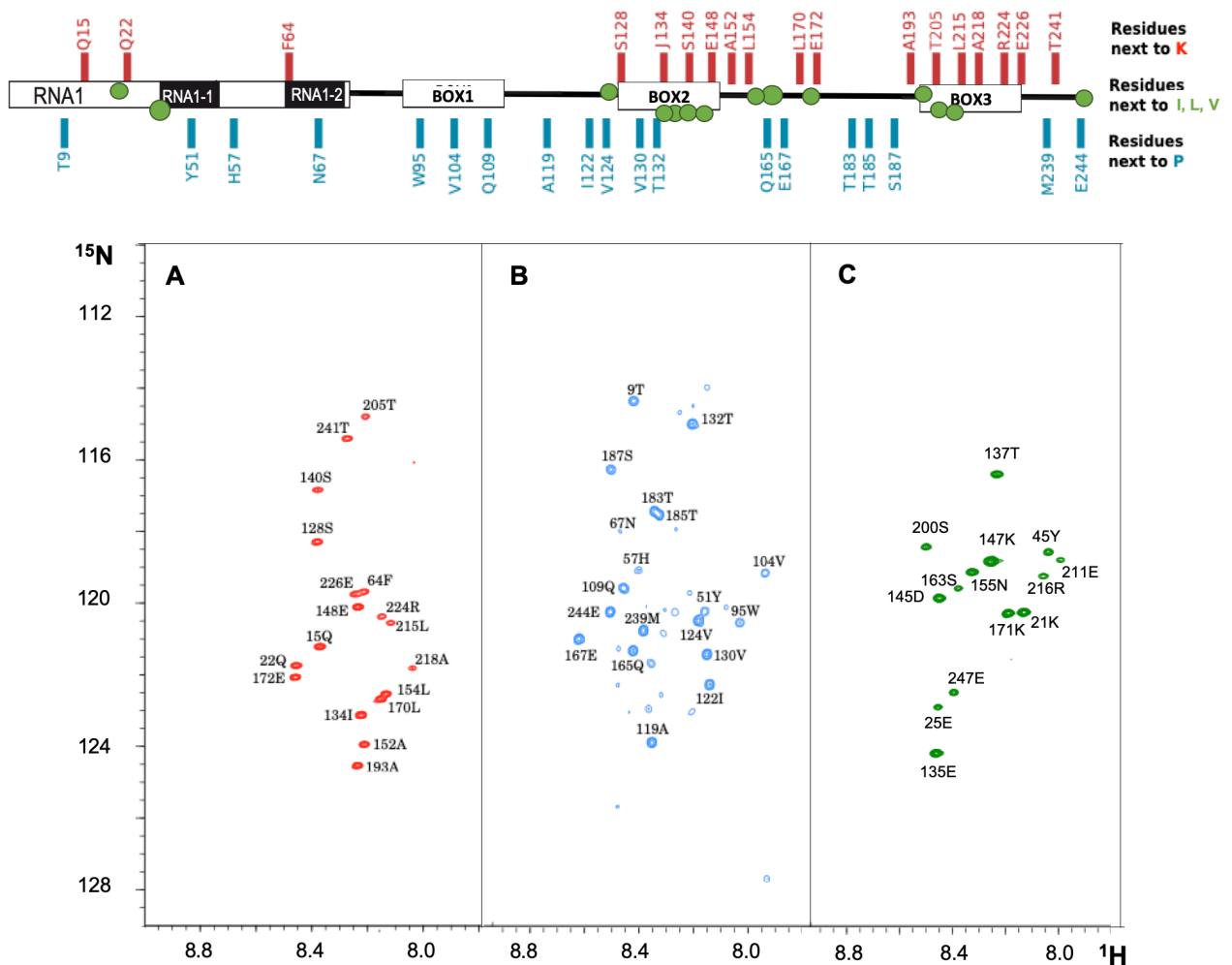

Figure 2. 2D ^{1}H-^{15}N HSQC experiment of eIF4G1 $_{1-250}$ reverse labelled samples of Lys(**A**), Pro(**B**) and I,L,V (**C**). The isotope discrimination pulse sequence selects only NH signals following these residues.

In the case of reverse labeling Lys, Pro or the α-ketoacid are added unlabeled to a $^{15}N/^{13}C$ minimum media. In fact, this approach not only can be used to simplify the spectra to facilitate the assign, nor can be used to study intramolecular interaction by paramagnetic relaxation enhancement (PRE), intermolecular self-recognition interactions (Chapter 6) or intermolecular protein-protein interactions (Chapter 7).

eIF4G1 $_{1-250}$ IDD contains residual structural features

2D ^1H-^{15}N HSQC NMR spectrum of eIF4G1 $_{1-250}$ (**Figure 3**) is characteristic of an IDP. Together with the major form (signals labelled in blue, **Figure 3**) we were able to identify several minor forms (signals labelled in red, **Figure 3**) that correspond to cisPro conformers and two uncommon posttranslational modifications at positions 41 and 76 that were assigned to isoaspartates (Schematic view of the chemical process of deamidation of aspartates, [20]). These variants arise from chemical isomerization of N^{41} and N^{76} which are part of NG sequences that have been reported to have the highest tendency to experience this non-enzymatic process in model peptides [33]. The level of deamidation is similar for the two positions (12 and 14 %) and was calculated integrating the S^{43} and G^{78} signals in both variants. The G^{42} and G^{77} deamidate isoforms overlap in the same site on the spectrum, so we could not choose them to estimate the population. Moreover, these level of deamidation remain constant between different samples and over NMR experimental time suggesting they were generated *in vivo* (by the bacteria). The analysis of eIF4G1 $_{1-250}$ ^{13}C chemical shifts, T_1/T_2 relaxation and residual dipolar couplings (RDCs) reveals structural and dynamic information. Chemical shifts are considered the most powerful tool for identifying secondary structure and furthermore, the study of RDCs is the tool used to corroborate the presence of secondary structure elements in IDPs or IDDs. The T_1/T_2 ratios inform about the rigidity or flexibility of along the sequence, and the presence of chemical exchange. The results of eIF4G1 $_{1-250}$ demonstrate the existence of a stable α-helix within BOX3 (**Figure 4A**), that was confirmed by characteristic sequential NOEs measured in a 3D ^{15}N-^{15}N-^1H HSQC-NOESY-HSQC spectra (**Figure S2A**). Its NMR structure was modeled using NOE data recorded from a BOX3 peptide (eIF4G $_{188-234}$) (**Figure S2B**). No further regular secondary structure elements were identified on eIF4G1 $_{1-250}$, but there are strong evidences of some source of residual high ordered structure present in the construct.

Figure 3. 2D ^{1}H-^{15}N HSQC spectrum of eIF4G1 $_{1-250}$ construct with labels for each assigned residues: major form (blue) and minor (red).
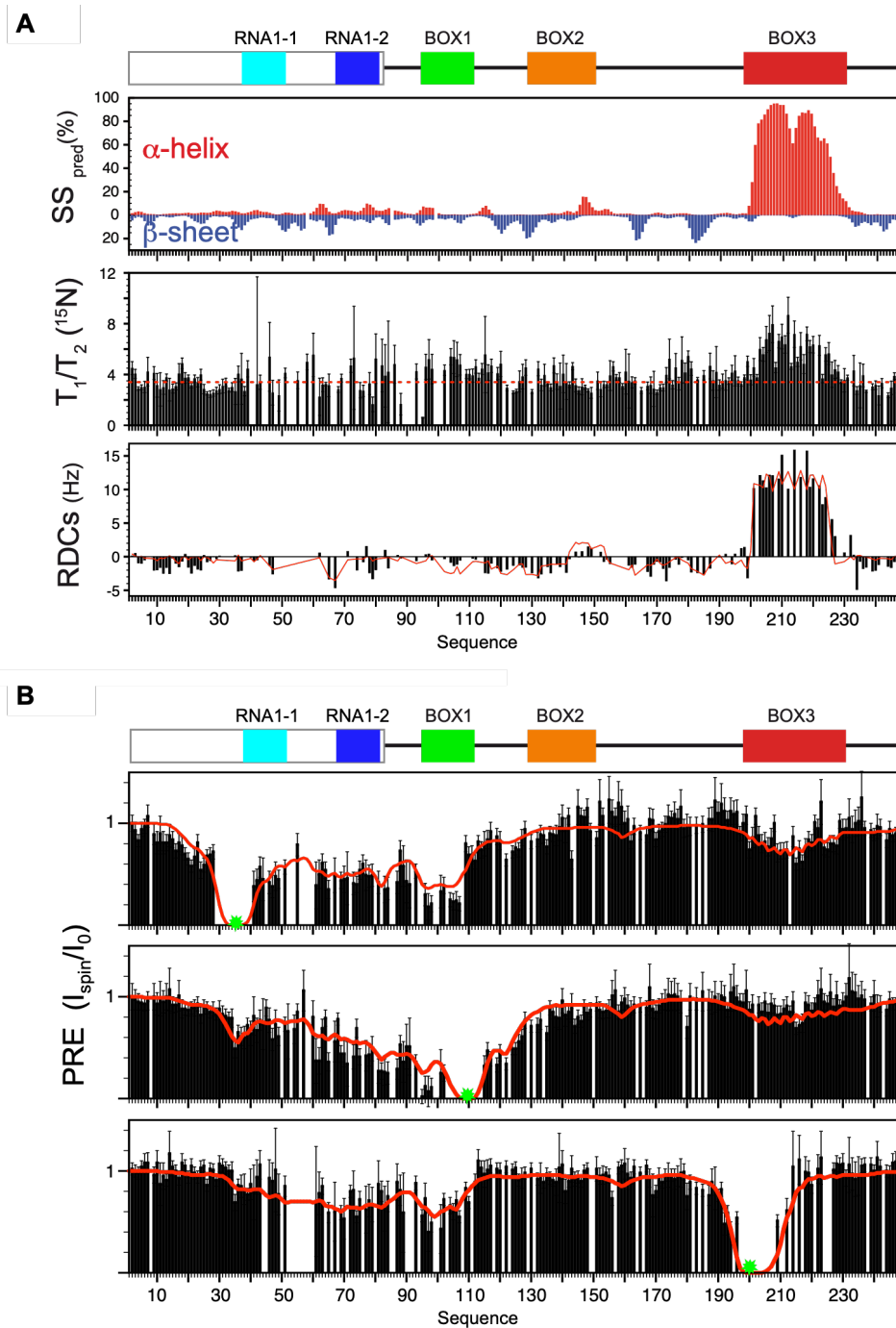
Figure 4. Structural analysis of eIF4G1 $_{1-250}$. (A) Evidences of residual secondary structure (per residue): percentage of secondary structure calculated with d2D (upper graph), ^{15}N relaxation T_1 over T_2 ratios (medium graph) and RDCs measured in Pf1 phage (lower graph). Conserved sequence elements have been represented over the histograms. **(B)** Per residue effects of Paramagnetic Relaxation Enhancements over relative signal intensities in three eIF4G1 $_{1-250}$ mutants: $G^{35}C$ (up), $Q^{109}C$ (medium) and $S^{200}C$ (low). Green spots mark the position of the mutations and red lines the averaged calculated values across the 158-ensemble of eIF4G1 $_{1-250}$ structure obtained in this work.

Several studies show broad a relatively disperse peaks in the NMR spectra (i.e. G^{97}, G^{77} and G^{83}) (**Figure 3**), not compatible with a fully disordered state. In addition the order parameters predicted from the chemical shifts [34], and the lower T_1/T_2 relaxation times for conserved boxes suggest that they might be involved in transient contacts that restrict mobility and/or induce chemical exchange contributions to S^2 (**Figure S3**).

But the definitive clue that support the existence of long-range interaction into eIF4G1 $_{1-250}$ are the paramagnetic relaxation enhancement (PRE) experiments. The nitroxyl spin labels at selected cysteine mutants (eIF4G1 $_{1-250}$ has no native CYS) unravels long-range PREs for $S^{200}C$, $Q^{109}C$ and $S^{35}C$ mutants (**Figure 4B**). The maximum range of PREs for eIF4G1 $_{1-250}$, calculated as described in [35], is expected to be 25 -30 Å (**Figure S4**). Therefore, the PRE data obtained for mutants, that have been strategically placed next to conserved boxes, reveal long-range contacts involving BOX1, RNA1 and BOX3 elements. We attempt other mutations within RNA1 element but result in constructs less stable and more prone to degradation. Similarly, the interactions involving BOX3 are presumably important for stability, as the construct lacking this element turn to be less stable (**Figure S1**).

Non-sequence specific cation/π and π/π interactions define eIF4G1 $_{1-250}$ structure

Defining the structure of IDPs or IDDs might well be seen as a semantic oxymoron. Rather than a single structure, the structures of IDPs are described by conformational ensembles (i.e. a collection of structurally-diverse conformers which on average best describe the properties of the IDP or IDD). To build these conformational ensembles is necessary to identify residual secondary structures and transient interaction between different regions of the polypeptide. The scarce nature of these transient contact makes it difficult to identify and quantify them by experimental methods because their lifetime and population are low, undermining the possibilities to calculate IDP structures. Alternatively, we could try to predict which residue-residue contacts might be probably to occur for a particular IDP based on other source of knowledge (e.g. biochemical data). Many studies have been made to understand the compositional bias of IDPs and to determine which residue types would favour aggregation, flexibility and/or long-range contacts. Along this line, a recent study shows that cation-π and π/π interactions are key for the "molecular grammar" of phase separation in prion-like IDPs [23].

We noticed that eIF4G1 $_{1-250}$ has residues capable for these interactions such as R[11] or Y[11], mostly located in conserved boxes (**Figure 1B**), and therefore we build conformation ensembles assuming the cation/π and π/π interactions dominant the long-range contact detected in the PRE data (**Figure 4B**). Structures were calculated using ambiguous cation/π distance restraints between Arg and Tyr, ambiguous π/π distance restraints between Tyr, Phe and Trp and other experimental restraints such as NOEs for the BOX3 α-helix. The final eIF4G1 $_{1-250}$ structure was selected from the original set of conformers (40000) using a protocol that builds the ensemble progressively, minimizing the deviations with the three PRE datasets (see Material and Methods). In this way we obtain a minimum conformation ensemble (158 members) that reproduce with very good agreement the PRE data (red lines in **Figure 4B**). The steps of the process used to build the conformational ensemble are schematically shown in **Figure S5**.

As expected, the structure of eIF4G $_{1-250}$ (**Figure 5A-B, Figure 6 and Figure S6**) is predominately disordered with the exception of BOX3 α-helix that also as the central element. The structure shows no predominant tertiary fold and is therefore structurally very diverse. The average contact map across the ensemble shows transient contacts among RNA1_1, RNA1_2 and BOX1 and between those and BOX3 (**Figure 5A** right). Interestingly, BOX2 appears no to be involved in this interaction networks and is more exposed (orange segments on **Figure 5A** left). When looking at the distribution of the contacts with the BOX3, we noticed that RNA1_1, RNA1_2 and BOX1 tend to occupy one side of the α-helix (**Figure 5B** right), which might bias the access to this structural element to other binding partners. The BOX2 in contrast is disperse and more evenly distributed around the central BOX3, in consistency with the lack of interactions.

The eIF4G1 $_{1-250}$ conformational ensemble shows a very high structural diversity within as can be seeing in the comparison of the contact maps of the individual structure (**Figure S6**). In general, the interactions between RNA1_1, RNA1_2, BOX1 and BOX3 seems to occur in networks involving these four elements rather than in binary mode, some examples are represented in **Figure 6**. Not surprisingly the major determinants of these networks are multiple π/π and cation/π contacts probably favoured by the planar nature of aromatic and guanidinium groups. In this way, each individual conformer shows a small protein core with unique potential for protein/ligand/RNA recognition.
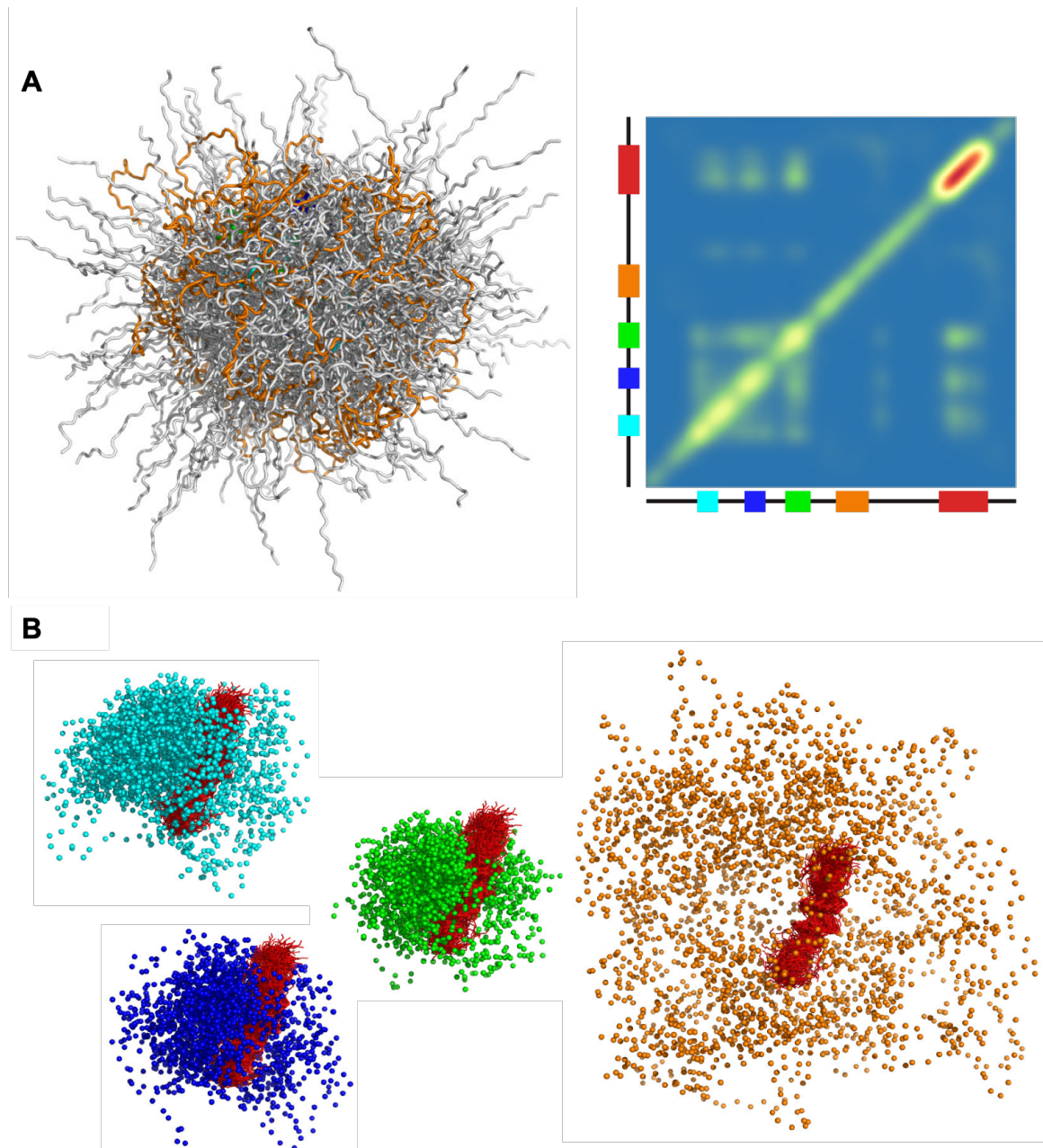
Figure 5. Structural analysis of eIF4G1 $_{1\text{-}250}$. (A) Superposition of the backbone traces of the 158-ensemble of eIF4G1 $_{1\text{-}250}$ structure. The C_α atoms of the α-helix in BOX3 have been used for alignment and this and conserved elements have been colored with the code used in previous panels. On the right, an ensemble-average C_α contact map showing the preferential long-range contacts between different parts of the eIF4G1 $_{1\text{-}250}$ (green/yellow/red). **(B)** The mapping of the C_α of RNA1_1, RNA1_2 and BOX1 (color-coded balls) for all the conformers of the eIF4G1 $_{1\text{-}250}$ ensemble reveals an slightly asymmetric population around BOX3 α-helix (red sticks). In contrast, BOX2 does not show any preferred interaction.
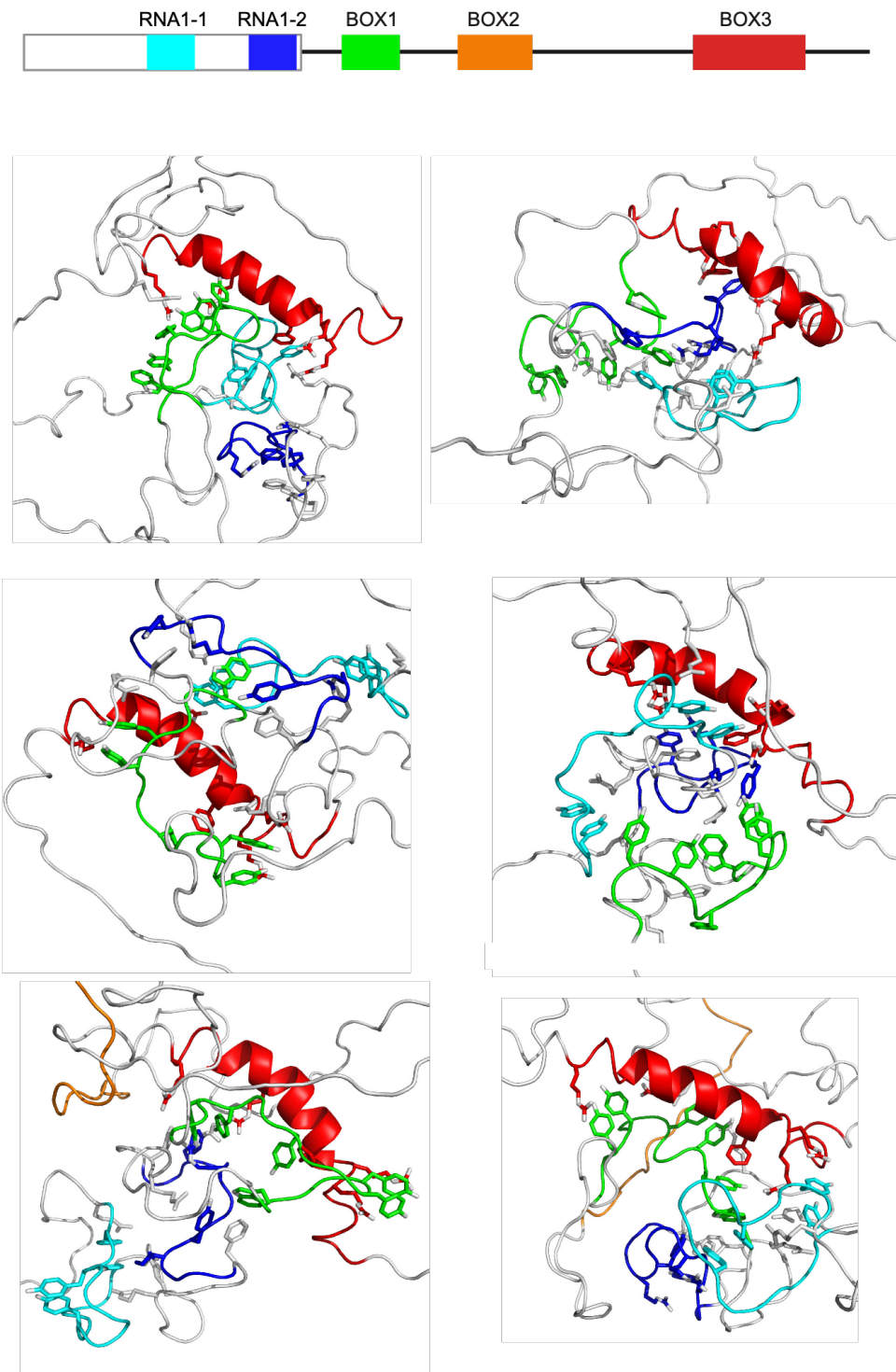
Figure 6. Structural analysis of eIF4G1 $_{1-250}$. Selected conformers of eIF4G1 $_{1-250}$ ensemble showing the side chains of Arg, Tyr, Phe and Trp. The examples illustrate the high structural diversity and dominance of extended π-π and π-cation networks. Conserved sequence elements have been represented over the conformers which have been colored with the code used in previous figures.

5.4. CONCLUSIONS

This work describes an extensive structural and dynamical characterization of eIF4G1 $_{1-250}$ at residue level using NMR. The NMR spectrum has been completely assigned, including minor forms, revealing two sites that suffer partial deamidation. The functional role of this chemical modification is still unknown, but it might be related with molecular aging processes as it happens in other proteins. The assignment of eIF4G1 $_{1-250}$ was greatly facilitated by the use of selectively labelling methods, exemplifying the great potential of these approaches to study large IDPs as eIF4G1 $_{1-250}$ and their interactions (shown in Chapter 7).

Parameters containing structural information in NMR such as RDCs, NOE data and PREs have been combined with a predicted knowledge-based interactions to calculate a large set of structures (40K) using a fast molecular dynamics protocol. Three PRE data sets have been used to build an ensemble eIF4G1 $_{1-250}$ that reproduce more faithfully the experimental restraints than other methods like Flexible-Meccano. The use of ambiguous π-cation and π-π interactions and the ensemble selection protocol, represent two novel strategies to obtain the structure of an IDP, perhaps one of the first examples of this kind.

The analysis of the ensemble shows concludes the existence of a highly populated α-helix within BOX3. Apart from that, the ensemble is conformationally diverse and the Arg and aromatic residues tend to group in networks rather than making binary isolated contacts. This result in a structure that is highly plastic, shows multiple mini-cores and a great degree of shapes with potential recognition surfaces. The eIF4G1 $_{1-250}$ ensemble is radically different of that of a well-folded structure and its characteristics are better described by statistical parameters like averaged contact maps. It does not resemble either the folding intermediates or structures like molten globules, were secondary structure elements are preformed and it is thought that there is a dominant (but not fixed) global fold. Instead the eIF4G1 $_{1-250}$ structure shows not a fully disordered state in which the conserved boxes RNA1_1, RNA1_2, BOX1 and BOX3 are involved in transient contacts, whereas BOX2 is not. These transient contacts would explain some of the behaviour of the construct, like for example the long-term stability which is rare in fully disordered proteins. Many of the transient contacts involve BOX3, which can be seen as a nucleation site due to the presence of stable helix.

The fact that constructs lacking BOX3 are less stable reinforce the idea that transient π-cation and π-π networks are important for stability, in eIF4G1 $_{1\text{-}250}$ and perhaps in other IDPs.

5.5.   MATERIALS AND METHODS

Cloning, protein expression and purification

Plasmids and proteins used in this work are described in the key resources table (**Appendix 3**). DNA fragments corresponding to wild-type construct of eIF4G1 $_{1\text{-}250}$ were amplified from *Saccharomyces cerevisiae* genomic DNA using DNA polymerase DNA polymerases KOD or Pfu. These DNA fragments were cloned in a pET28-modified vector that contains: N-terminal thioredoxin A fusion tag, an internal 6xHis tag and a TEV protease site. The mutants in eIF4G1 $_{1\text{-}250}$ ($S^{200}C$, $Q^{109}C$ and $S^{35}C$) were obtained with a Quick-change Lightning Kit and specific DNA primers. Plasmids were transformed in *E. coli* BL21 (DE3) competent cells and expressed in kanamycin containing (30 µg/l) LB medium.

For isotopic labelling, a K-MOPS derived minimal medium [36] was supplemented with $^{15}NH_4Cl$ (1 g/l) and/or ^{13}C-glucose (4 g/l). In case of amino acid-selective isotope labeling, *Escherichia Coli* autotrophic strains RF6 and RF10 has been used for production of Pro or Lys reverse-labeled eIF4G$_{1\text{-}250}$ samples. The K-MOPS minimal medium was supplemented with $^{15}NH_4Cl$ (1 g/l) and/or ^{13}C-glucose (4 g/l) and Proline (0.5 g/l) or Lysine (0.125 g/l), depends of the autotrophic strain used [28]. Cultures of eIF4G1 $_{1\text{-}250}$ and mutants were grown at 37 ºC until OD_{600} nm = 0.6 – 0.8 when were induced with 0.5 µM of IPTG during 4 hours. For all proteins, cell pellets were resuspended in lysis buffer (25 mM Potassium phosphate pH 8.0, 300 mM NaCl, 10 mM Imidazole and 1 tablet/50 ml of protease inhibitors cocktail), lysed by sonication and cleared by ultracentrifugation.

The supernatant was purified by metal affinity chromatography using HiTrap^{TM} 5ml column and eluted using the buffer 25 mM Potassium phosphate pH 8.0, 300 mM NaCl and 300 mM Imidazole. The samples with the fusion protein were exchange to 20 mM Tris pH 8.0 and digested overnight at 4 ºC with homemade TEV protease. Then, we purify eIF4G1 $_{1\text{-}250}$ and mutants away from uncleaved, txA and TEV using a cation exchange column (SP 5ml). Finally proteins were concentrated and buffer exchanged according to their posterior use.

NMR experiments

All samples were prepared in NMR buffer (25 mM Phosphate pH 6.5, 25 mM NaCl, 1mM DTT and 10 % D_2O) and experiments acquired at 25 ºC on cryoprobe-equipped Bruker AV800 MHz spectrometer. The assignment of the backbone ^{1}H, ^{15}N and ^{13}C atoms was achieved by following the standard methodology. The 3D experiments 3D HNCO, 3D HNCA, 3D HC(CO)CA [24], 3D HN(CA)CO [25], 3D CBCANH [26], 3D CBCA(CO)NH [27] and 3D (H)CCH-TOCSY which were recorded to assign the side chain resonances [37]. Proteins concentration range between 100 – 200 µM. The chemical shifts of eIF4G1 $_{1-250}$ will be deposited in the Biomagnetic Resonance Database (BMRB). Measurements of ^{15}N backbone amide relaxation T_1 and T_2 were measured with standard inversion-recovery and CPMG spin echo series of ^{1}H-^{15}N spectra. NMR spectra were processed using TOPSPIN v2.1 (Bruker, Inc) and NMRPipe [38], and they were analyzed with the program CcpNmr Analysis v2.4.2 [39].

Residual dipolar Couplings (RDCs)

We used two types of anisotropic media to induce weak alignment of eIF4G1 $_{1-250}$: alcohol mixtures and Pf1 phages [40]. For the first ones were composed we prepared 5 % $C_{12}E_5$/Hex r (0.85) by adding 3 µl of hexanol, and 5 % $C_{12}E_6$/Hex r (0.64) with 4.4 µl of hexanol to 25 mM Phosphate pH 6.5 and 25 mM NaCl containing the protein at 200 µM concentration. A ^{2}H quadrupolar splitting of 30 Hz was observed for $C_{12}E_5$ media and appeared to stabilize faster than the $C_{12}E_6$. The filamentous phages Pf1 were used a final concentration of 20 mg/ml rendering a ^{2}H quadrupolar splitting of 8.45 Hz. The protein (200 µM) is dissolved in 25 mM Phosphate pH 6.5 and 150 mM NaCl. The NMR experiments were carried out at 298 K on a Bruker Avance III 800 MHz spectrometer equipped with a cryogenic triple resonance probe. For the measurements of RDCs (D), two samples (isotropic and anisotropic) have to be prepared in order to measure the coupling in both conditions. Indeed, the isotropic sample gives access to the scalar J-coupling, and the anisotropic one to the addition of the scalar and bipolar (J+D) couplings.

These measurements have been done for both alignment media (5% $C_{12}E_5$/Hex r(0.85) and filamentous phage Pf1) with the ^{1}H-^{15}N-HSQC-DSSE (In Phase Anti Phase IPAP) for both the oriented and the isotropic samples. Stability and integrity of the sample was monitored by regular 2D ^{1}H-^{15}N-HSQC

experiments in order to rule out interactions between the protein and the alignment media. Experiments were processed using TOPSPIN v2.1 (Bruker, Inc) and NMRPipe [38], and they were analyzed with the program CcpNmr Analysis v2.4.2 [39].

Nitroxyl spin labelling and Paramagnetic Relaxation Enhancement (PRE) measurements

Protein samples from different cysteine containing eIF4G $_{1-250}$ mutants ($S^{200}C$, $Q^{109}C$ and $S^{35}C$) were chemically modified with the following protocol. Between 600 – 700 µM of the samples were exhaustively reduced with 5 mM DTT for two hours at room temperature. The DTT was eliminated by fast buffer exchange to 25 mM Tris pH 9.0 and 25 mM NaCl using desalting column. Labeling with 4-(2-Iodoacetamido)-TEMPO was initiated immediately after column elution by adding a tenfold molar excess of the spin label dissolved in ethanol (25 mM spin label stock). Reaction was lead to proceed for 30 minutes at room temperature at darkness.

The excess of 4-(2-Iodoacetamido)-TEMPO was quenched with 10 mM of 2-mercaptoethanol for 10 minutes and afterwards the protein adduct was exchange to 25 mM Phosphate pH 6.5, 25 mM NaCl and 1 mM DTT for later use. The NMR samples were prepared in 5 mm sealed tubes in a N_2 atmosphere to avoid oxidation by air and high resolution 2D ^1H-^{15}N HSQC were recorded for the reduced state (active spin label). After that, the spin label was oxidized with 10 µM ascorbate [41] to record the reference 2D ^1H-^{15}N HSQC without paramagnetic relaxation enhancement. The relaxation effect was calculated as the intensity rations between peaks in the two spectra.

Structure calculation of eIF4G1 $_{188-250}$ peptide.

The NMR structure of the peptide eIF4G1 $_{188-250}$ was determined from NOE-derived distance restrains (2D NOESY spectra with 60 ms mixing time) and angular restrains. The data was processed and analyzed using the program Cyana 3.0 [42]. Samples were obtained by recombinant expression and purified as previously described for eIF4G1 $_{1-250}$. eIF4G1 $_{188-250}$ assignments was obtained by comparison with eIF4G1 $_{1-250}$ constructs (^{15}N-^{15}N planes in the 3D ^{15}N-HSQC-NOESY-^{15}N-HSQC spectra of eIF4G1 $_{1-250}$ show the characteristics sequential HN-HN NOEs between consecutive residues), and confirmed by triple resonance 3D spectra: CBCA(CO)NH, HNCACB and HNCO.

Structure calculation of eIF4G1 $_{1-250}$

The first set of eIF4G1 $_{1-250}$ ensembles were generated with Flexible-Meccano [43]. Random conformers (70000) were generated with conformational propensities defined by the analysis of chemical shift data with the program CamShift [34]. Then $^1D_{HN}$ RDCs values were back calculated for the conformers, averaged across the whole ensemble and compared with experimental ones [40]. The process was repeated interactively by modifying the populations until no further improve of the agreement between the experimental and simulated RDC values. The we used 40000 of these conformers as starting structures for a second protocol with the program Cyana 3.0 [42] that follows standard restricted molecular dynamics protocol. For each conformer a simulated annealing was performed using experimental and knowledge-based restraints. As experimental restrains we used NOE data for the BOX3 peptide and a 35 Å upper distance limits for all the PRE ratios below 0.7.

This limit has been generously chosen according the theoretical relationship, for eIF4G1 $_{1-250}$, between signal intensity ratios and distance. The knowledge-based restraints were implemented considering the recent publication that highlight the importance of π-cation and π-π interactions in IDPs [23]. Ambiguous π-cation restraints were set up between Arg and Phe/Tyr and, similarly, ambiguous π-π pairwise interactions between the aromatics (Phe,Tyr and Trp). In this way no biases of specific pairwise iterations were introduced.

eIF4G1 $_{1-250}$ ensemble built up.

Theoretical PRE-derived intensity ratios were calculated for each of the 40000 conformers using the distance between the HNs and the position of the spin label on each of the three experimental sets (S^{200}, Q^{109} and S^{35} positions) and the **equations 1 − 4** [35] (**Figure S4**). These theoretical ratios were compared with experimental ones and the deviations (r^2) added to obtain a global value T_{PRE} . The conformer with minimum T_{PRE} was chosen as the ensemble seed. In the second interaction the T_{PRE} values were calculated as averaged between the values r^2 of the seed and each of the remaining structures in the dataset. Once again the pair structure leading to a minimum value of the T_{PRE} was selected. The protocol was followed iteratively until the last structure and T_{PRE} represented as a function of the ensemble size (**Figure S5**).

We determine the size ensemble as from a T_{PRE} value of 10 % above from the minimum (158 members). All the calculations were performed with home-made Perl scripts. The theoretical values of the PREs for the three conditions and other structural properties, were calculated as averages across this ensemble.

Equation 1. The intensity ratio can be expressed as a function of the relaxation rates with and without spin label and the INEPT evolution time (10 ms).

$$\frac{I_{ox}}{I_{red}} = \frac{R_2 e^{-(R_2^{spin} t)}}{R_2 + R_2^{spin}}$$

Equation 2. The proton relaxation rate calculated from the width at average height in the HSQC.

$$R_2 = \pi \Delta \nu_{1/2}$$

Equation 3. Relaxation with the spin label.

$$R_2^{spin} = \left[\frac{K}{d^6} \left(4\tau_c + \frac{3\tau_c}{1 + \omega_I^2 \tau_c^2} \right) \right]$$

Equation 4. The correlation time determined by ^{15}N relaxation from the T_1/T_2 ratio.

$$\tau_c = \frac{1}{4\pi \Delta \nu_N} \sqrt{6 \frac{T_1^N}{T_2^N} - 7}$$

5.6.   ACKNOWLEDGMENTS

5.7. REFERENCES

1.   Das S, Das B. eIF4G—an integrator of mRNA metabolism? FEMS Yeast Res. 2016 Oct 14;16(7). Available from: https://doi.org/10.1093/femsyr/fow087

2.   Hentze MW. eIF4G: A Multipurpose Ribosome Adapter? Science (80). 1997 Jan 24;275(5299):500 LP – 501. Available from: http://science.sciencemag.org/content/275/5299/500.abstract

3.   Hershey JWB, Sonenberg N, Mathews MB. Principles of Translational Control: An Overview. 2012;1–10.

4.   Sonenberg N, Hinnebusch AG. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell. 2009/02/26. 2009;136(4):731–45. Available from: https://www.ncbi.nlm.nih.gov/pubmed /19239892

5.   Byrd MP, Zamora M, Lloyd RE. Generation of Multiple Isoforms of Eukaryotic Translation Initiation Factor 4GI by Use of Alternate Translation Initiation Codons. 2002;22(13):4499–511.

6.   Clarkson BK, Gilbert W V, Doudna JA. Functional Overlap between eIF4G Isoforms in Saccharomyces cerevisiae. 2010;5(2).

7.   Berset C, Zurbriggen A, Djafarzadeh S, Altmann M, Trachsel H. RNA-binding activity of translation initiation factor eIF4G1 from Saccharomyces cerevisiae. RNA. 2003 Jul;9(7):871–80. Available from: https://www.ncbi.nlm.nih.gov/pubmed/12810920

8.   Imataka H, Gradi A, Sonenberg N. A newly identified N-terminal amino acid sequence of human eIF4G binds poly(A)-binding protein and functions in poly(A)-dependent translation. EMBO J. 1998;17(24):7480–9.

9.   Santiveri CM, Mirassou Y, Rico-Lastres P, Martinez-Lumbreras S, Perez-Canadillas JM. Pub1p C-terminal RRM domain interacts with Tif4631p through a conserved region neighbouring the Pab1p binding site. PLoS One. 2011/09/21. 2011;6(9):e24481. Available from: https://www.ncbi.nlm.nih.gov/pubmed /21931728

10.   Buchan JR, Muhlrad D, Parker R. P bodies promote stress granule assembly in Saccharomyces cerevisiae. J Cell Biol. 2008/11/05. 2008;183(3):441–55. Available from: https://www.ncbi.nlm.nih.gov/ pubmed/18981231

11.   Buchan JR, Yoon JH, Parker R. Stress-specific composition, assembly and kinetics of stress granules in Saccharomyces cerevisiae. J Cell Sci. 2010/12/22. 2011;124(Pt 2):228–39. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21172806

12.   Hoyle NP, Castelli LM, Campbell SG, Holmes LE, Ashe MP. Stress-dependent relocalization of translationally primed mRNPs to cytoplasmic granules that are kinetically and spatially distinct from P-bodies. J Cell Biol. 2007/10/03. 2007;179(1):65–74. Available from: https://www.ncbi.nlm.nih.gov /pubmed/17908917

13.   Goyer C, Altmann M, Lee HANS, Blanc A, Deshmukh M, Woolford JL, et al. TIF4631 and TIF4632 : Two Yeast Genes Encoding the High- Molecular-Weight Subunits of the Cap-Binding Protein Complex (Eukaryotic Initiation Factor 4F) Contain an RNA Recognition Motif-Like Sequence and Carry Out an Essential Function. 1993;13(8):4860–74.

14. Tarun Jr. SZ, Sachs AB. Association of the yeast poly(A) tail binding protein with translation initiation factor eIF-4G. EMBO J. 1996/12/16. 1996;15(24):7168–77. Available from: https://www.ncbi.nlm.nih.gov/pubmed/9003792

15. Haar T Von Der, Mccarthy JEG. Intracellular translation initiation factor levels in Saccharomyces cerevisiae and their role in cap-complex function. 2002;531–44.

16. Chemical C, Plant R, Plant P, Biol M, Ghaemmaghami S, Huh W, et al. Global analysis of protein expression in yeast. 2003;108(1997):737–41.

17. Kessler SH, Sachs AB. RNA recognition motif 2 of yeast Pab1p is required for its functional interaction with eukaryotic translation initiation factor 4G. Mol Cell Biol. 1998;18(1):51–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9418852%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC121449

18. Park EH, Walker SE, Lee JM, Rothenburg S, Lorsch JR, Hinnebusch AG. Multiple elements in the eIF4G1 N-terminus promote assembly of eIF4G1*PABP mRNPs in vivo. EMBO J. 2010/12/09. 2011;30(2):302–16. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21139564

19. Park E-H, Walker SE, Lee JM, Rothenburg S, Lorsch JR, Hinnebusch AG. Multiple elements in the eIF4G1 N-terminus promote assembly of eIF4G1•PABP mRNPs in vivo. EMBO J. 2011;30(2):302–16. Available from: http://dx.doi.org/10.1038/emboj.2010.312

20. Martinez-Lumbreras S. The role of Gbp2p , Nab2p and Pub1p along the mRNA cycle : structural and molecular recognition studies by NMR and other biophysical techniques (Doctoral thesis). Univ Autónoma Madrid. 2013;

21. Rasia RM, Brutscher B, Plevin MJ. Selective isotopic unlabeling of proteins using metabolic precursors: application to NMR assignment of intrinsically disordered proteins. Chembiochem . 2012;13(5):732–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22408059

22. Golovanov AP, Blankley RT, Avis JM, Bermel W. Isotopically discriminated NMR spectroscopy: A tool for investigating complex protein interactions in vitro. J Am Chem Soc. 2007;129(20):6528–35.

23. Wang J, Choi JM, Holehouse AS, Lee HO, Zhang X, Jahnel M, et al. A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. Cell . 2018/07/03. 2018;174(3):688-699 e16. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29961577

24. Grzesiek S, Bax A. Improved 3D triple-resonance NMR techniques applied to a 31 kDa protein. J Magn Reson. 1992;96(2):432–40.

25. Clubb RT, Thanabal V, Wagner G. A constant-time three-dimensional triple-resonance pulse scheme to correlate intraresidue 1HN, 15N, and 13C' chemical shifts in 15N13C-labelled proteins. J Magn Reson. 1992;97(1):213–7.

26. Grzesiek S, Bax A. An efficient experiment for sequential backbone assignment of medium-sized isotopically enriched proteins. J Magn Reson. 1992;99(1):201–7.

27. Grzesiekt S, Bax A. Correlating Backbone Amide and Side Chain Resonances in Larger Proteins by Multiple Relayed Triple Resonance NMR. J Am Chem Soc. 1992;114(16):6291–3.

28. Lin MT, Fukazawa R, Miyajima-nakano Y, Matsushita S, Choi SK, Iwasaki T, et al. Escherichia coli Auxotroph Host Strains for Amino Acid-Selective Isotope Labeling of Recombinant Proteins. 1st ed. Vol. 565, Isotope labeling of biomolecules labeling Methods. Elsevier Inc.; 2015. 45–66 p. Available from: http://dx.doi.org/10.1016/bs.mie.2015.05.012

29. Verardi R, Traaseth NJ, Masterson LR, Vostrikov V V, Veglia G. Isotope Labeling for Solution and Solid-State NMR Spectroscopy of Membrane Proteins Polarization Inversion Spin Exchange at Magic Angle. 2012;35–62.

30. Zerbs S, Frank AM, Collart FR. Bacterial Systems for Production of Heterologous Proteins . 1st ed. Vol. 463, Methods in Enzymology. Elsevier Inc.; 2009. 149–168 p. Available from: http://dx.doi.org/10.1016/S0076-6879(09)63012-3

31. Grady CO, Rempel BL, Sokaribo A, Nokhrin S, Dmitriev OY. One-step amino acid selective isotope labeling of proteins in prototrophic Escherichia coli strains. Anal Biochem. 2012;426(2):126–8. Available from: http://dx.doi.org/10.1016/j.ab.2012.04.019

32. Waugh D. Genetic tools for selective labeling of proteins with α-15N-amino acids. J Biomol NMR. 1996;8(2):184–92.

33. Robinson NE, Robinson AB. Prediction of protein deamidation rates from primary and three-dimensional structure. Proc Natl Acad Sci U S A. 2001/04/11. 2001;98(8):4367–72. Available from: https://www.ncbi.nlm.nih.gov/pubmed/11296285

34. Camilloni C, De Simone A, Vranken WF, Vendruscolo M. Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. Biochemistry. 2012/03/01. 2012;51(11):2224–31. Available from: https://www.ncbi.nlm.nih.gov/ pubmed/22360139

35. Battiste JL, Wagner G. Utilization of site-directed spin labeling and high-resolution heteronuclear nuclear magnetic resonance for global fold determination of large proteins with limited nuclear overhauser effect data. Biochemistry. 2000/05/23. 2000;39(18):5355–65. Available from: https://www.ncbi.nlm.nih.gov/pubmed/10820006

36. Neidhardt FC, Bloch PL, Smith DF. Culture medium for enterobacteria. J Bacteriol. 1974;119(3):736–47.

37. Sattler M, Schleucher J, Griesinger C. Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. Prog Nucl Magn Reson Spectrosc. 1999;34(2):93–158.

38. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. NMRPipe: A multidimensional spectral processing system based on UNIX pipes. J Biomol NMR. 1995;6(3):277–93. Available from: http://www.scopus.com/inward/record.url?eid=2s2.00029400480&partnerID=40&md5=2ca830f73de65c42691425cd22747e94

39. Skinner SP, Fogh RH, Boucher W, Ragan TJ, Mureddu LG, Vuister GW. CcpNmr AnalysisAssign : a flexible platform for integrated NMR analysis. J Biomol NMR. 2016;66(2):111–24.

40.   Almond A, Axelsen JB. Physical Interpretation of Residual Dipolar Couplings in Neutral Aligned Media. 2002;9986–7.

41.   Gillespie JR, Shortle D. Characterization of long-range structure in the denatured state of staphylococcal nuclease. I. Paramagnetic relaxation enhancement by nitroxide spin labels. J Mol Biol. 1997;268(1):158–69.

42.   Güntert P. Automated NMR Structure Calculation With CYANA. In: Downing AK, editor. Protein NMR Techniques. Totowa, NJ: Humana Press; 2004. p. 353–78. Available from: https://doi.org/10.1385/1-59259-809-9:353

43.   Ozenne V, Bauer F, Salmon L, Huang JR, Jensen MR, Segard S, et al. Flexible-meccano: A tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. Bioinformatics. 2012;28(11):1463–70.

5.8. SUPPORTING INFORMATION
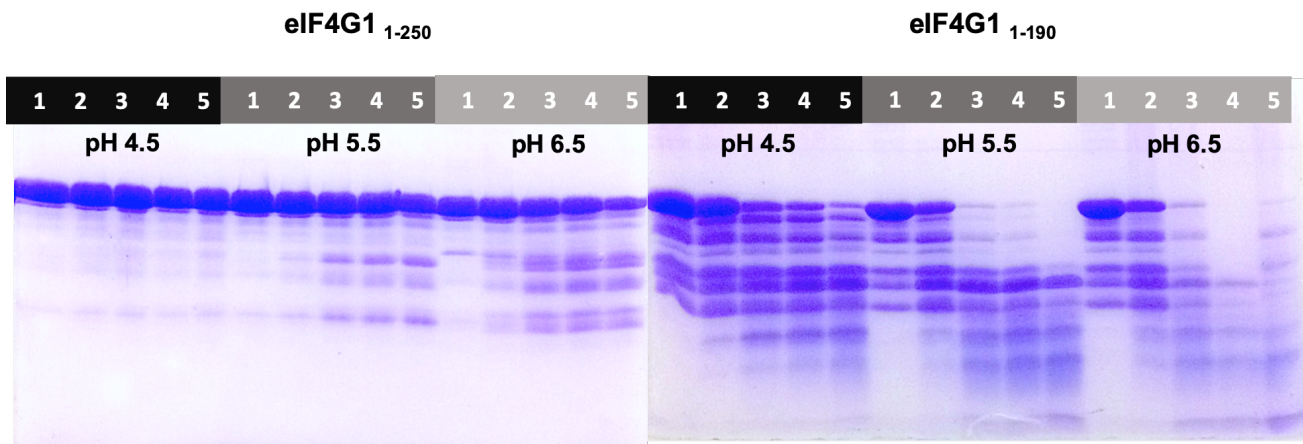
Figure S1. Compared stability study of 50 μM eIF4G1 $_{1\text{-}250}$ and eIF4G1 $_{1\text{-}190}$ under different pH conditions. We compared the pH effect in five time points: 1 (0 h), 2 (24 h), 3 (4 days), 4 (5 days) and 5 (8 days). conditions used were 25 mM acetate or phosphate buffer, 25 mM NaCl and 0.0005 % of Azide
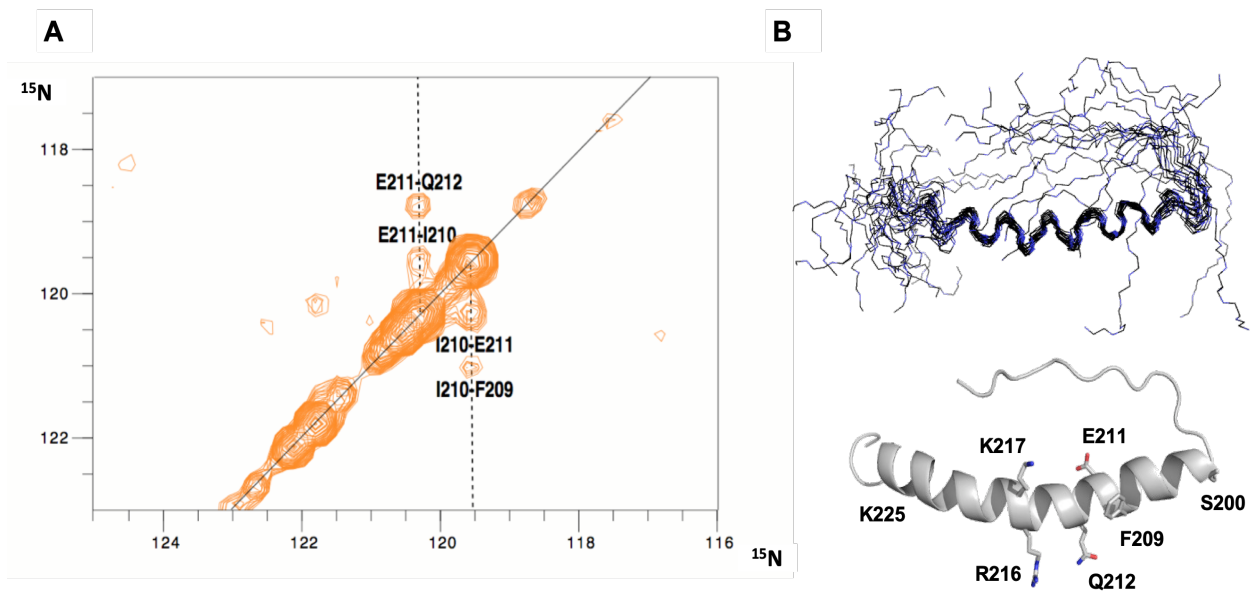
Figure S2. (A) The ^{15}N-^{15}N planes in the 3D ^{15}N-HSQC-NOESY-^{15}N-HSQC spectra of eIF4G1 $_{1-250}$ show the characteristics sequential HN-HN NOEs between consecutive residues of the α-helix in BOX3. **(B)** The NMR structure of the BOX3 peptide forms a continuous α-helix 7-turns long with the conserved residues placed in the middle section oriented towards two sides of the helix.
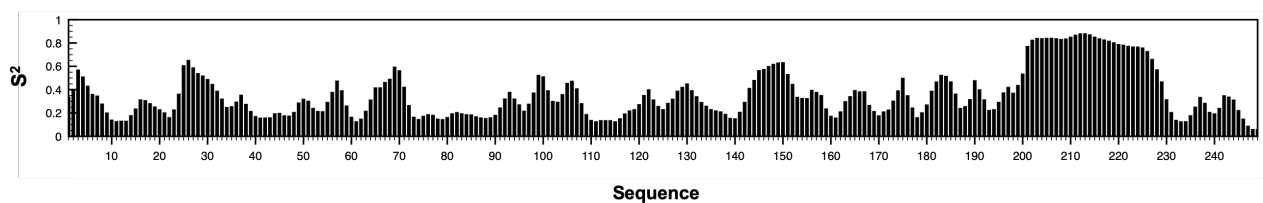
Figure S3. S^2 sequence distribution of eIF4G1 $_{1-250}$ predicted from chemical shifts using the program Camshift. S^2 values from eIF4G1 $_{1-250}$ BOX3 predicts more rigidity comparing with the rest of domains.
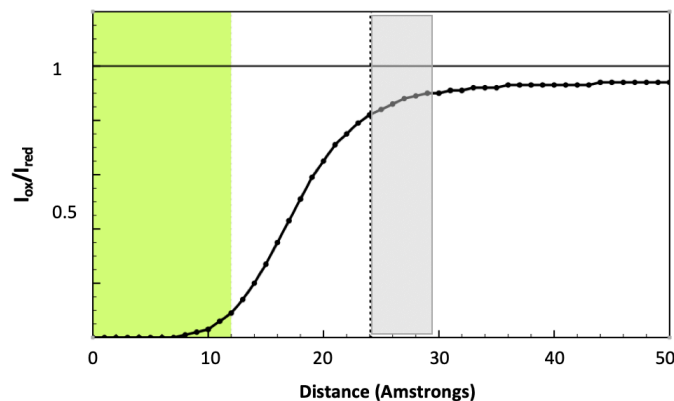
Figure S4. Theoretical PRE-derived intensity ratios (I_{ox}/I_{red}) use the distance between the HNs and the position of the spin label on each sample. In eIF4G1 case we use the spin label in S^{200}, Q^{109} and S^{35} positions and the maximum range of PREs calculated is expected to be 25 -30 Å (grey area).

Figure S5. Scheme of the process steps used to build the representative conformational ensemble (158 structures, green circle) of eIF4G1 $_{1-250}$.
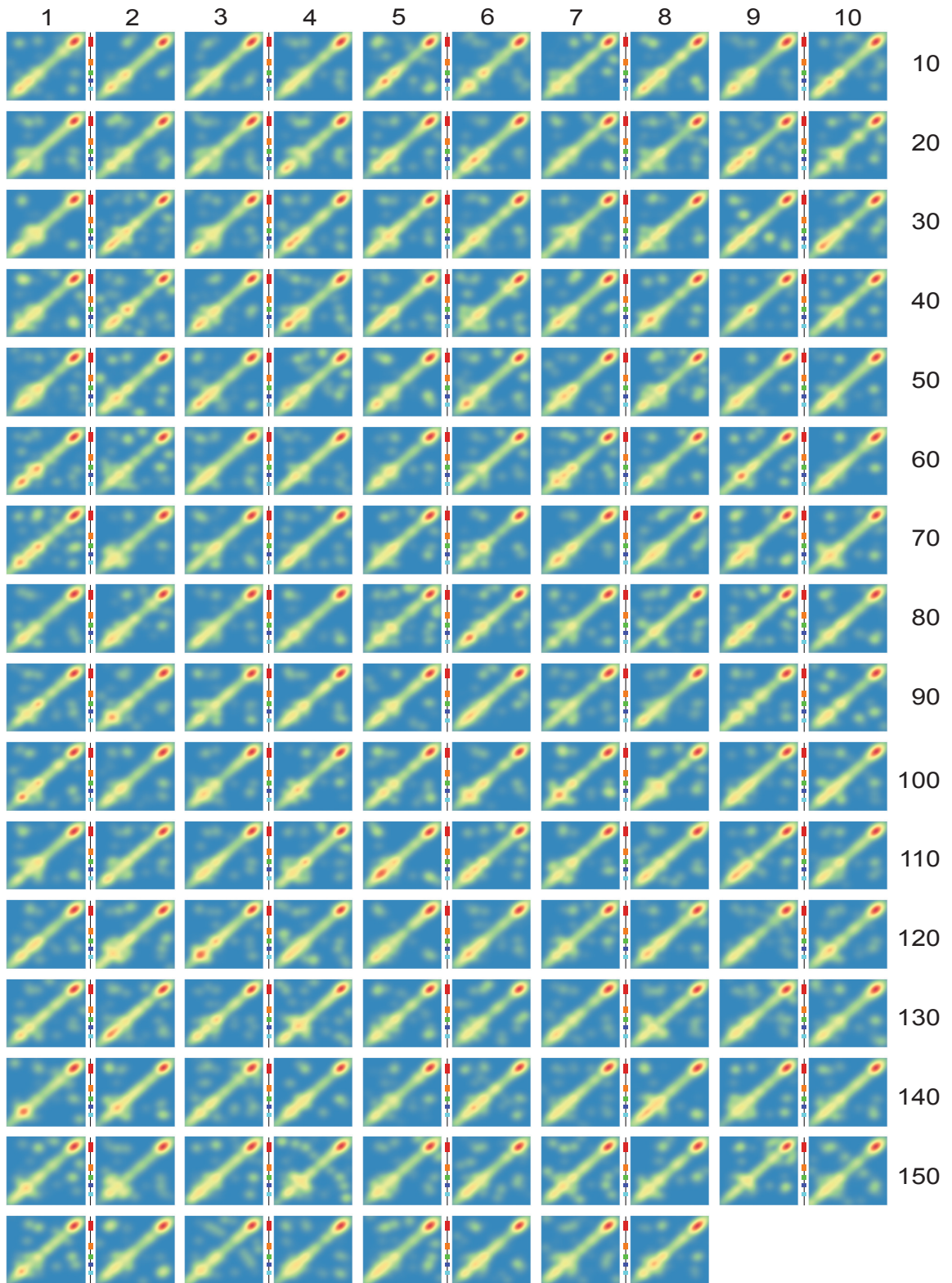
Figure S6. Contact maps for each of the 158 conformers of eIF4G1 1-250 conformational ensemble. Conserved sequence elements have been represented with the code used in previous figures.

CHAPTER 6

Self-recognition properties of eIF4G1 N-terminal IDP and RNA binding proteins Pub1 and Pab1

6.1. ABSTRACT

RNA-binding proteins (RBPs) form large macromolecular assemblies with RNA that govern essential biological processes. They typically have a multidomain structure with (multiple) small RNA-binding domains (e.g. RRM, KH, zinc fingers, etc) and large low complexity domain (LCD) predicted as intrinsically unstructured and often containing repeated sequences (e.g. RGG boxes). Both RBDs and LCDs can bind RNA and protein and many times self-recognize. This characteristic domain architecture enables multivalence and can potentially form large molecular networks.

Poly(U) binding protein (Pub1) and poly(U) binding protein Pab1 are two highly abundant RNA-binding proteins in *Saccharomyces cerevisiae* that interact with eIF4G1 and are crucial for translation control. This chapter describes the NMR study of Pab1 and Pub1 constructs that include RBDs and LCDs, particularly focusing on their self-recognition properties. The results, together with the study of eIF4G1 self-association, is an important previous step to understand the cooperative role of these three proteins in molecular processes such as translation initiation or of stress granules nucleation.

This study shows that the intrinsically disordered protein (IDP) eIF4G1 $_{1-250}$ can dimerize/oligomerize through contacts among BOX1 and RNA1 regions. On the other hand, regarding the RBPs, we found that Pub1 and Pab1 can also form specific dimers and circular species in which participate one or two RRM domains depending on the protein, and in some cases segments of the LCDs. The study is based in a combination of nuclear magnetic resonance (NMR) and SEC-MALS and highlights the mechanistic importance of self-recognition in combination with heterologous interactions between the three proteins (eIF4G1 $_{1-250}$, Pub1 and Pab1) that will be studied in detail on the next chapter.

6.2. INTRODUCTION

RNA binding proteins play essential roles in mRNA metabolism. In *Saccharomyces cerevisiae*, poly(A)-binding protein 1 or Pab1 [1,2] and poly(U)-binding protein 1 (Pub1) are among the most abundant mRNA binding proteins [3], Pab1 and Pub1 contain four and three RRMs respectively, which typically function as (i) RNA recognition motifs, but can also serve in (ii) protein-protein recognition. The two proteins have various low complexity regions (LCD) that have been involved in liquid-liquid phase separation or pre-nucleation of stress granules [4]. The structural domains of these three proteins and their conserved sequences are represented in **Figure 1**. Pub1 and Pab1 interact with eIF4G1 N-terminal IDD [5] using RRM domains, and through various sequences within the eIF4G1 $_{1-250}$ (as will be shown in Chapter 7).

Pab1 is involved in many intracellular functions associated with mRNA metabolism: deadenylation, mRNA nuclear export, translation initiation and termination [6]. Until recently there were not crystallography or NMR structures for yeast Pab1 RRMs by , but a reliable models, derived from its human homolog [7,8], can be constructed using folding prediction servers (i.e. Swiss-model https://swissmodel.expasy.org). However a very recent electron microscopy study has revealed the structural details of the poly(A) RNP in complex with the deadenilation machinery [9]. The experimental models for Pab1 RRMs are very similar to the predicted ones, thus we continue using these one on this thesis. In contrast, there is a vast amount of biochemical and functional studies for Pab1 and it is commonly admitted that thanks to its modular architecture Pab1 can simultaneously interact with poly(A) tails and different proteins that regulate the translation process. This protein plays an important role in translation because it promotes the assembly of ribosomal subunits onto the mRNAs, and is presence is required by different mechanisms to translate mRNAs into proteins [6]. Furthermore, Pab1 could influence gene expression under stressful condition by affecting the formation of quinary assemblies. These structures are known as transient protein assemblies that are kept together by multivalent intermolecular interactions [10], which are formed in response to specific stress conditions [11]. Pab1 undergoes LLPS and its LCD have been involved on this physical phenomena [12]. On the other hand, some studies have propose other mechanism of dimerization/oligomerization, which involve the RRMs [13].

Some of them has been recently proved structurally [12]. This Chapter focused in the description of this second type of specific self-assembling mechanism, seeking to define which regions are involved on the process. The other protein of this study, Pub1, is specifically involved in the stability and translational control of many mRNAs, [14]. Several Pub1 structures has been solved by [5,15]. The RRM1 and RRM2 domains of Pub1 bind poly(U) RNA with high selectivity [5,16], whereas the C-terminal RRM domain (RRM3) of Pub1 binds U-rich and AU-rich sequences, is responsible for the interaction with eIF4G1 and contains an a non-canonical N-terminal helix [5]. Pub1 biological function antagonizes the Pab1 principal one, in translation initiation. The evidences supporting this are (i) Pub1 seems to take part of the translationally inactive mRNAs pool [16], and (ii) is one of the first described stress granules components in budding yeast [17]. As for Pab1, recent studies showed that Pub1 is capable to LLPS, a process for which RRMs seem to be critical whereas the LCD plays a more accessory role. [18]. The goal of this chapter is to extend the structural organization of multidomain proteins Pab1 and Pub1 and to study their self-association in greater detail. More specifically, this study is based on the characterization of the self-recognition of Pub1, Pab1 and, as well as eIF4G1 $_{1-250}$. In the end, the new findings would be essential together with the information in the other chapters to propose a unified model for these three proteins and their role in gene expression regulation.
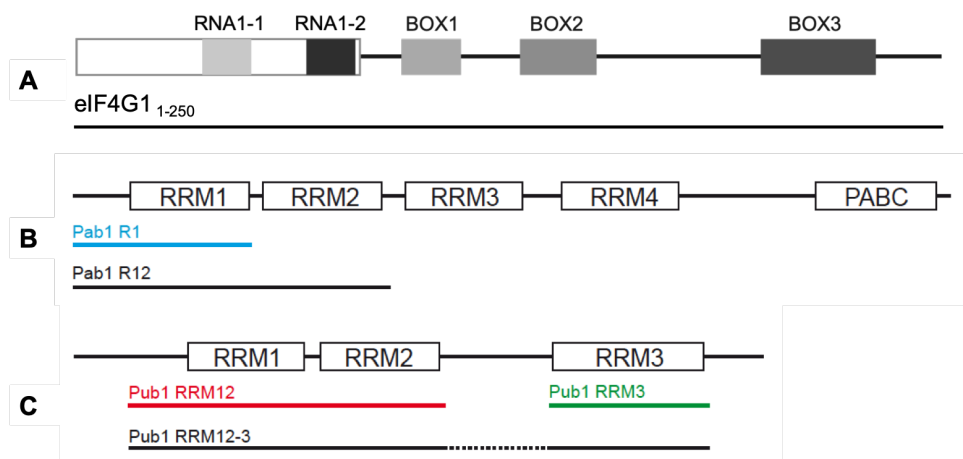
Figure 1. Schematic representation of **(A)** *S. Cerevisiae* eIF4G1 $_{1-250}$ containing the conserved sequence boxes (RNA1_1, RNA1_2, BOX1, BOX2 and BOX3), **(B)** *S. Cerevisiae* Pab1 RRM domains (RRM1, 2, 3, and 4), C-terminal domain (PABC) and Pab1 constructions used in this work (Pab1 RRM1 in blue and Pab1 RRM12 in black), and **(C)** *S. Cerevisiae* Pub1 RRM domains (RRM1, 2 and 3) and the Pub1 construction are represent in black (Pub1 RRM123), red (Pub1 RRM12) and green (Pub1 RRM3).

6.3. RESULTS AND DISCUSSION

eIF4G1 $_{1-250}$ IDD self-recognition

The intramolecular interactions between RNA1_1, RNA1_2, BOX1 and BOX3 of eIF4G1 $_{1-250}$ shown in Chapter 5 (Q^{109}C example in **Figure 2A**), can potentially occur intermolecularly. To analyse this possibility, we performed an additional Paramagnetic Relaxation Enhancement (PRE) experiment placing the spin label in *trans*: attaching it to a Q^{109}C mutant (unlabelled) and monitoring the effects in wild-type eIF4G1 $_{1-250}$ ^{15}N-labelled.
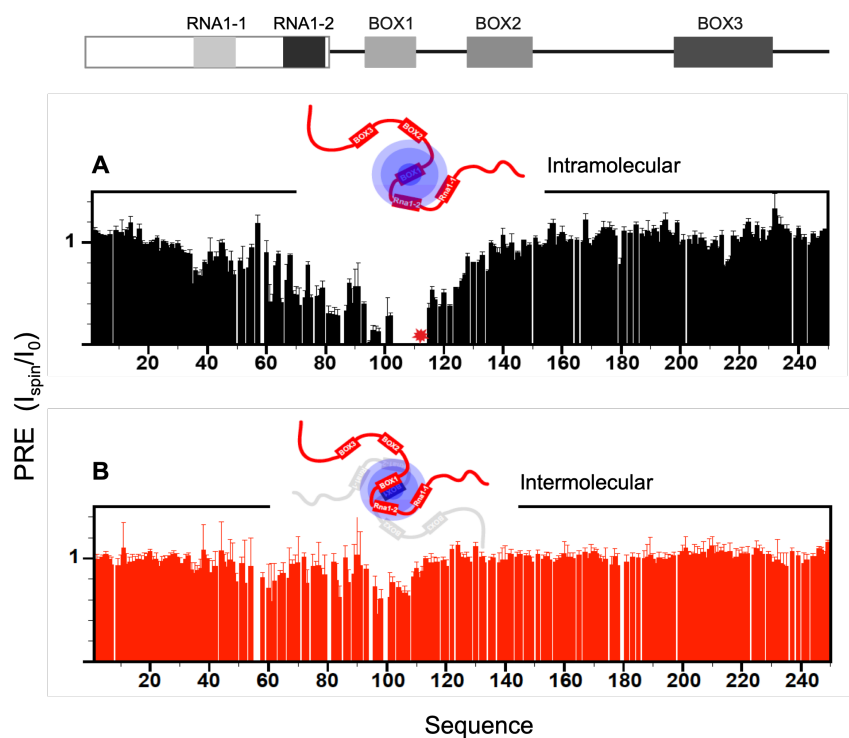
Figure 2. Comparison of PRE effects for the eIF4G1 $_{1-250}$ Q^{109}C mutant with 4-(2-Iodoacetamido)-TEMPO spin label in cis (A, black histogram) and in trans (B, red histogram). Intermolecular contacts affect RNA and BOX1 elements preferentially, but the intramolecular interactions dominate in the free eIF4G1 $_{1-250}$.

The results (**Figure 2B**) show that eIF4G1 $_{1-250}$ can dimerize/oligomerize through contacts involving BOX1 and RNA1 regions, as these elements "sense" the presence of the spin label in the mutant molecules. However, the comparison with the Q^{109}C PRE data suggest that dimers/oligomers remain at low population in the free state and therefore the intramolecular contacts dominate.

The intermolecular PREs does not seems to affect BOX3 but rather involve interactions of BOX1 with elements to its N-terminus or itself. In this context, is interesting to notice that BOX1 contains a predicted pro-amyloidogenic sequence (ZIPPERDB, [19]), it is plausible that the intermolecular PREs are actually monitoring early stage species (i.e. dimers) of this oligomerization process.

Pub1 and Pab1 self-recognition

Pub1 and Pab1 have a similar architecture with 3 and 4 RRM domains respectively (**Figure 1B-C**) and various LCD. These domains, specially the RRMs, that have been related to the LLPS phenomena as we mentioned in the Introduction section. Because it has been suggested that RRMs self-association can be an important factor of LLPS, this study we focus in protein-protein self-recognition through these domains. For that, (i) we studied the protein dynamics behaviour by measuring ^{15}N relaxation parameters (T_1 and T_2), and (ii) we studied the intermolecular interaction using 2D ^{1}H-^{15}N HSQC NMR experiments, in which we could analysed perturbations on the signal intensity and/or chemical shift. The results analysed let us to know the interactions between different parts of Pub1 and Pab1.

^{15}N relaxation data for Pub1 RRM12 and Pub1 RRM3 constructs show that the RRMs behave as dynamically independent folded domains while the N-terminal part of Pub1 RRM12 does as an IDP (**Figure S1**). However, in the 2D ^{1}H-^{15}N HSQC spectrum of Pub1 RRM123 (**Figure 3A**, in black), a construct that combine both constructions (Pub1 RRM12 and Pub1 RRM3), shows a dramatic reduction in the signal to noise. This suggests that the RRMs arrange in a compact form, rather than a bead-on-string model. Therefore, we studied if Pub1 RRM12 and RRM3 constructs can interact, by comparing their individual spectra with that of mixture (**Figure 3A**, in red and green with that of a black ones).

The signal intensities ratios (**Figure 3B**) reveal that Pub1 RRM3 interacts transiently with Pub1 RRM12 thought discrete regions in the RRMs and, more importantly, with a short hydrophobic segment ($V_{60}VPANAI_{66}$) within the Pub1 N-terminal IDP region (**Figure 3B**). This dimerization mechanism is specific and probably enhanced in the Pub1 RRM123 protein because it could occur intramolecularly in this case.
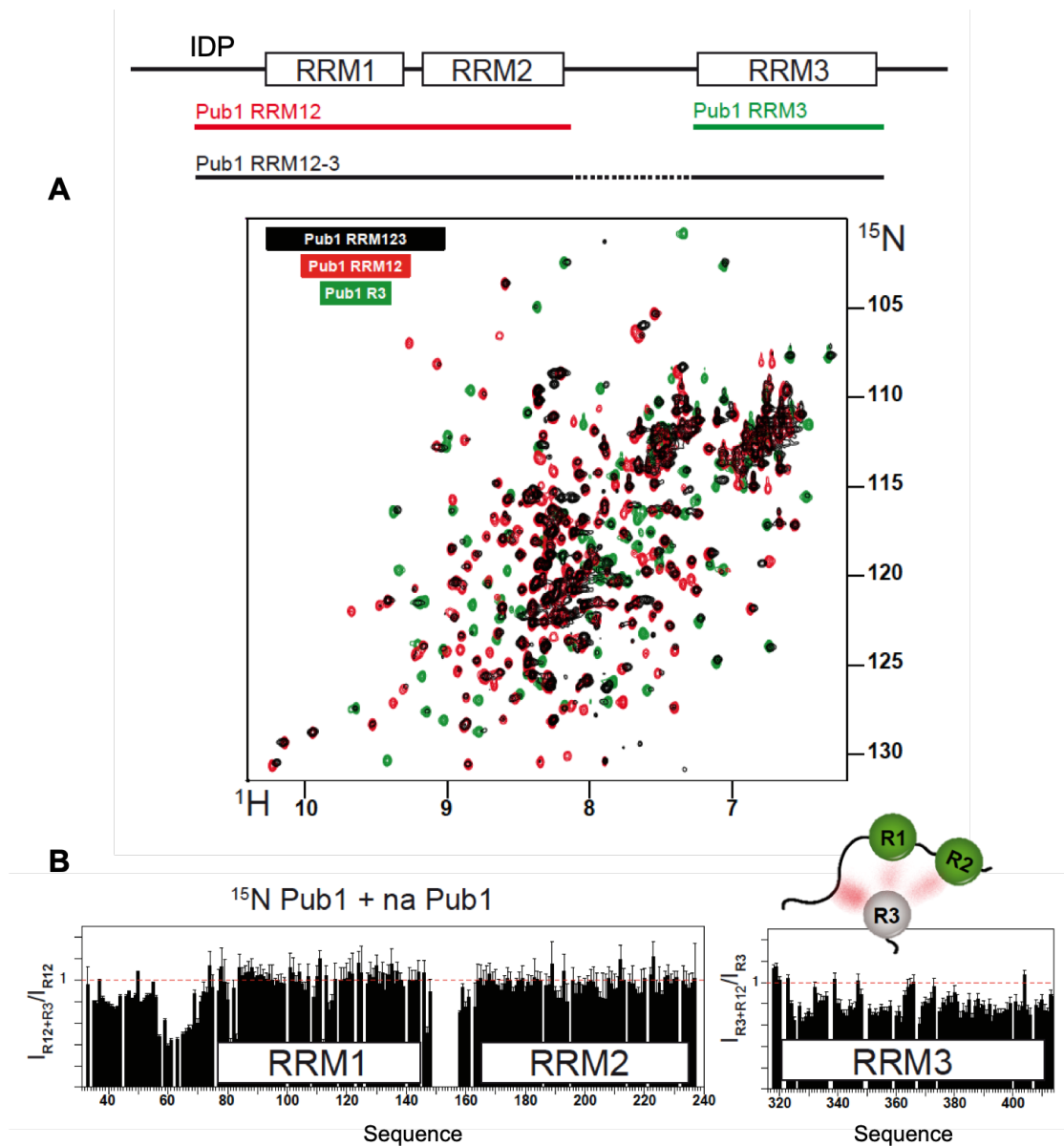
Figure 3. **(A)** 2D ^{1}H-^{15}N HSQC experiment of Pub1 RRM12 (red), Pub1 RRM3 (green) and Pub1 RRM123 (black). **(B)** Intensity ratios of ^{15}N Pub1 RRM12 and natural abundance (na) Pub1 RRM3 (left histogram), and na Pub1 RRM12 with ^{15}N Pub1 RRM3 (right histogram). The schematic model of the intermolecular interaction is represented in green balls (Pub1 RRM12) and grey ball (Pub1 RRM3).

Interestingly, a similar mechanism involving intramolecular circular species and intermolecular dimers (or oligomers) has been proposed for yeast Pab1 [20]. We assigned the 2D ^{1}H-^{15}N HQSC spectrum of the Pab1 RRM12 construct (**Figure 4A**). Despite Pab1 RRM12 has a similar architecture/size than Pub1 RRM12, the ^{15}N relaxation T_1/T_2 ratios are larger (**Figure S1**), suggesting that RRMs mobility is more

restricted in this case. Moreover, the spectra broaden progressively with the increase of concentration, proving that Pab1 RRM12 self-associates. To identify the elements involved in this self-association, we performed NMR titration experiments by mixing unlabelled Pab1 RRM12 over a fixed amount of ^{15}N labelled Pab1 RRM12, at equivalent ratio and at two different concentrations (**Figure 4B**).
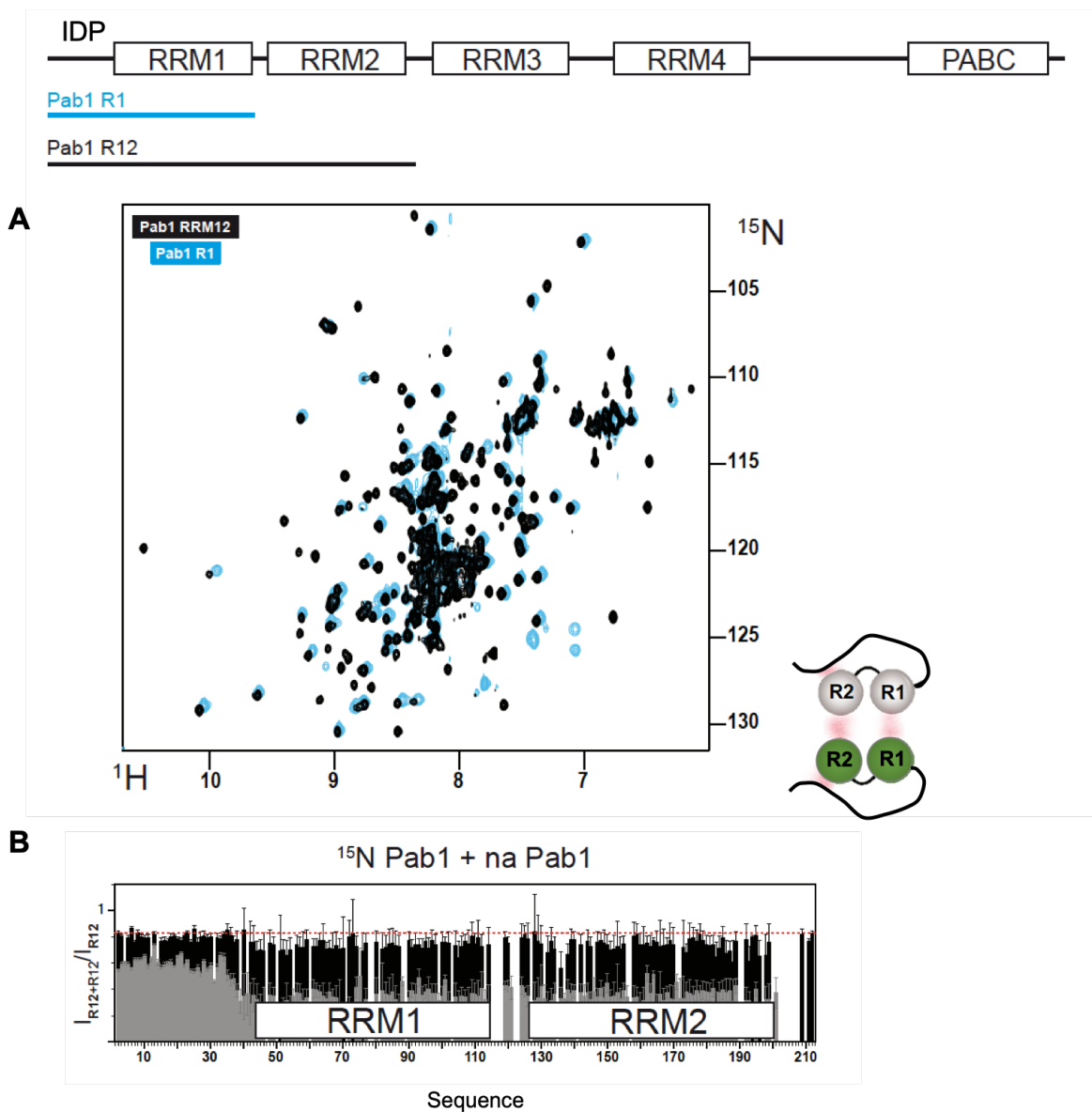
Figure 4. **(A)** 2D ^{1}H-^{15}N HSQC experiment of Pab1 RRM12 (black) and Pab1 RRM1 (blue). **(B)** Intensity ratios of ^{15}N Pab1 RRM12 and natural abundance (na) Pab1 RRM12 at 100 μM (black histogram) compared with equivalent data but with the proteins at 290 μM (grey histogram). The schematic model of the intermolecular interaction is represented in green balls (Pab1 RRM12) and grey balls (Pab1 RRM12).

The results show a selective decrease of intensities for RRM1 and RRM2 signals (compared to the Pab1 IDP) which becomes more evident at higher concentration. A similar effect occurs for Pab1 RRM1 showing that this domain alone can self-associate and that dimerization mechanism involves RRM1-RRM1 contacts (**Figure S2**). Furthermore, the analysis of the intensity ratios of Pab1 RRM1 samples at different concentrations also suggests the RRM1 self-association (**Figure S3**). We could not determine if isolated Pab1 RRM2 has the same trait, because it is very low solubility makes its NMR study inaccessible. Finally, we compared the spectra of Pab1 RRM1 and RRM12 recorded at equal concentration and buffer conditions, to look for the changes induced by RRM2 in RRM1 (**Figure S4**). The chemical shift perturbations are small and discrete and the intensities are reduced in the RRM region and also for residues 10 – 20 (EQLENLNIQDD), a leucine-rich region in the IPD part of the construct that contains Pab1 Nuclear Export Signal [21]. The later evidences the existence of transient contacts between the IDP and RRM2 proposed in the schematic model of **Figure 4**.

In addition, we used SEC-MALS to study Pub1 and Pab1 dimerization. The results of Pub1 RRM123 in which coexist a monomer-dimer equilibrium, are represented in **Figure 5A**. Based on this, and NMR data, we propose an interaction model in which Pub1 RRM3 interact with the N-terminal Pub1 IDP (an some minor contacts with RRM1 and 2) to form a ring-like structure (monomer) or an elongated form (dimmer). On the other hand, the SEC-MALS data of Pab1 confirms also a self-association, in this case likely involving contacts between both RRMs and between the NES and RRM2. (**Figure 5B**). In the **Table 1** are represented all data that confirm the dimerization of both proteins. Interestingly, we did not observe signs of interaction between Pab1 RRM12 and Pub1 RRM12 and RRM3 constructs suggesting that the RRM-RRM and IPD-RRM interactions are protein-specific (self-recognition) and Pab1 and Pub1 do not form heterodimers. Previous structural studies exemplify the great diversity of RRM homodimerization modes. These domains can use multiple modes to self-associate: through face-to-face β-sheet contacts like Nup35 [22]; through side-to-side β-sheet contacts like in PAB1N [23] or through helix-helix contacts like in RPBMS [24,25] and HuC RRM3 [26]. The present data on Pub1 and Pab1 do not allow to say which of these binding modes actually occurs. Pup1, and to lesser extend Pab1, self-associates through heterologous interfaces (i.e. contacts between different RRMs), and through IPD-RRM interactions which not previously addressed in detail.

Table 1. SEC-MALS data. $M_w/M_n = M_z/M_n$: identify holds for a long-normal distribution where M_n is the number-average molar mass, M_w is the weight-average molar mass and M_z is the z-average molar mass.

	PROPERTY	PUB RRM 123		PAB RRM 12	
		Monomer	Dimer	Monomer	Dimer
Polydispersity	M_w/M_n	1.001 (1%)	1.083 (2%)	1.014 (6%)	1.073 (19%)
	M_z/M_n	1.003 (2%)	1.155 (4%)	1.027 (9%)	1.161 (36%)
Molar Mass Moments (g mol^{-1})	M_n	3.301e+4 (0.9%)	5.539e+4 (2%)	2.118e+4 (5%)	3.659e+4 (12%)
	M_w	3.306e+4 (0.9%)	6.002e+4 (2%)	2.147e+4 (4%)	3.926e+4 (14%)
	M_z	3.311e+4 (2%)	6.396e+4 (3%)	2.174e+4 (8%)	4.247e+4 (34%)
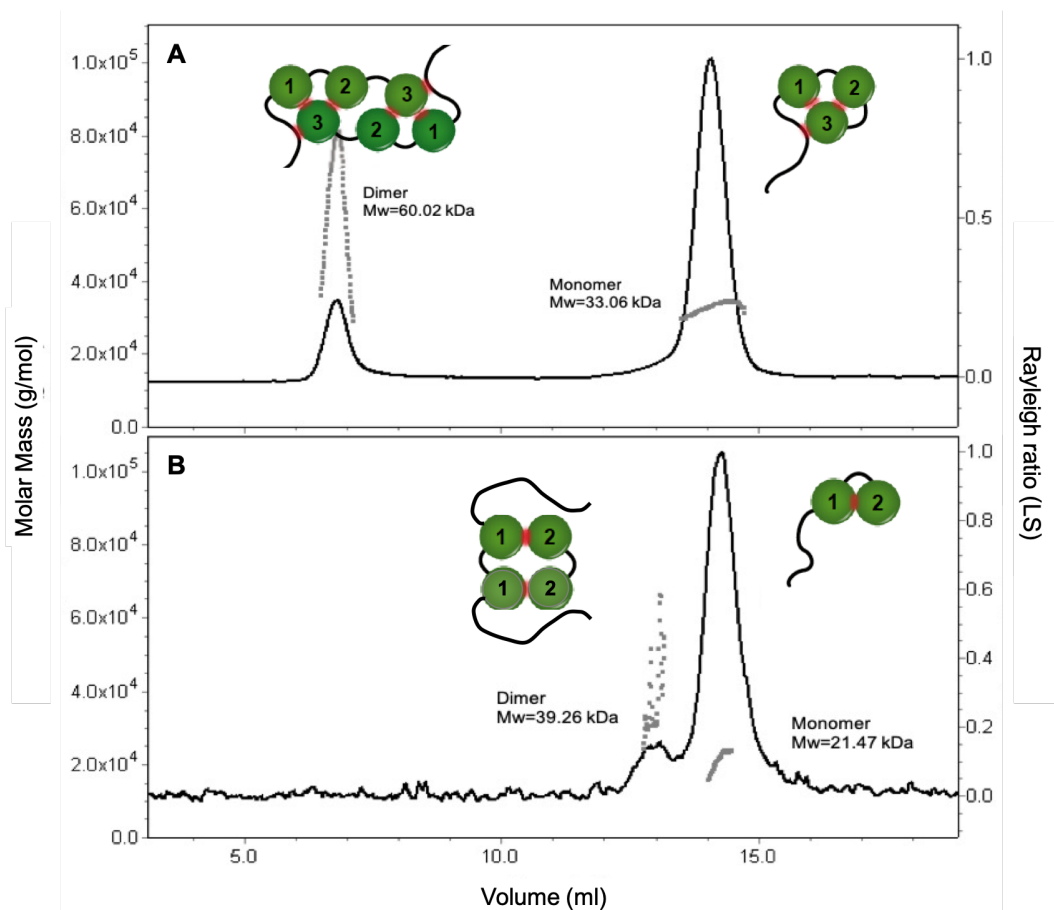
Figure 5. SEC-MALS elution volume for Pub1 RRM123 (A) and Pab1 RRM12 (B) RNA binding proteins. Signals form the 90º angle light-scattering detector or LS (black full line) and molar mass (grey squares) versus elution volume plot are shown in each case.

6.4. CONCLUSIONS

This chapter analyses the dimerization and, in general, self-recognition properties of eIF4G1 $_{1-250}$, Pub1 and Pab1. These types of interactions are important for multivalence, and in the case of multidomain proteins, like Pab1 and Pub1, or IDPs showing complex conformational behaviour, like eIF4G1 $_{1-250}$, can raise to extensive oligomeric networks and ultimately to LLPS. Since the three proteins have been reported to experience LLPS, it is important to understand in more detail the structural basis of self-recognition. This chapter deals with this objective as a previous step to analyse the heterologous interactions between the three proteins (Chapter 7). Using intermolecular PRE, this study demonstrates that eIF4G1 $_{1-250}$ can dimerize or oligomerize through contacts involving a specific region of this sequence (BOX1 and RNA1 domains). It is hypothesized that the nature of contacts is similar to those found in Chapter 5: π-π and π-cation interactions involving aromatic and arginine residues. However the extend of the intermolecular effects is smaller than intramolecular ones, concluding that dimers or oligomers are present at low population in the free state.

Regarding the RNA binding proteins, this study concludes that Pub1 and Pab1 can also form dimers and circular species involving specific interactions between RRMs and IDP domains. For Pub1 the key interactions involve RRM3 and a short hydrophobic segment in the N-terminal IDP. This self-recognition properties could be important for Pub1 LLPS as recently reported and this work provides a precise definition of the regions/domains involved. In the case of Pab1 RRM12 the two RRMs appears to be dynamically coupled probably by transient contacts between the RRMs and RRM1 is involved in dimerization, as was previously suggested [20]. This study discovers another transient contact involving the NES, in the N-terminus, and RRM2 that might have implications on its nucleo-cytoplasm trafficking.

6.5. MATERIALS AND METHODS

Cloning, protein expression and purification

Plasmids and proteins used in this work are described in the key resources table (**Appendix 3**). DNA fragments corresponding to wild-type constructs of eIF4G1 $_{1-250}$, Pub1 and Pab1 were amplified from *Saccharomyces cerevisiae* genomic DNA using DNA polymerase DNA polymerases KOD or Pfu.

These DNA fragments were cloned in a pET28-modified vector that contains: N-terminal thioredoxin A fusion tag, an internal 6xHis tag and a TEV protease site. The mutant in eIF4G1 $_{1-250}$ (Q^{109}C) was obtained with a Quick-change Lightning Kit and specific DNA primers. Plasmids were transformed in *E. coli* BL21 (DE3) competent cells and expressed in kanamycin containing (30 µg/l) LB medium. For isotopic labelling, a K-MOPS derived minimal medium [27] was supplemented with ^{15}NH$_4$Cl (1 g/l) and/or ^{13}C-glucose (4 g/l). In case of amino acid-selective isotope labeling, *Escherichia Coli* autotrophic strains RF6 and RF10 has been used for production of Pro or Lys reverse-labeled eIF4G$_{1-250}$ samples. The K-MOPS minimal medium was supplemented with ^{15}NH$_4$Cl (1 g/l) and/or ^{13}C-glucose (4 g/l) and Proline (0.5 g/l) or Lysine (0.125 g/l), depends of the autotrophic strain used [28]. Cultures of eIF4G1$_{1-250}$ were grown at 37 ºC until OD$_{600}$ nm = 0.6 – 0.8 when were induced with 0.5 µM of IPTG during 4 hours. Pab1 and Pub1 cultures, after reaching OD$_{600}$ nm = 0.6, were transferred to 25 ºC to induced with IPTG overnight (12 – 16 hours).For all recombinant proteins, cell pellets were resuspended in lysis buffer (25 mM Potassium phosphate pH 8.0, 300 mM NaCl, 10 mM Imidazole and 1 tablet/50 ml of protease inhibitors cocktail), lysed by sonication and cleared by ultracentrifugation. The supernatant was purified by metal affinity chromatography using HiTrap^{TM} 5ml column and eluted using the buffer 25mM Potassium phosphate pH 8.0, 300 mM NaCl and 300 mM Imidazole. The samples with the fusion protein were exchange to 20 mM Tris pH 8.0 (in case of Pab1 construct this buffer was supplemented with 1 mM DTT), and digested overnight at 4 ºC with homemade TEV protease.

In case of Pub1 and Pab1 constructs, the samples were reloaded on the HiTrap nickel column to capture the protease, cleaved fusion protein and uncleaved. The flow through was further purified by ion exchange using an anion exchanger column (Q 5ml) for all the proteins except Pub1 RRM3 that was purified with a cation exchange column (SP 5ml). In either case, proteins were eluted with a linear salt

gradient (to 1 M NaCl). In the case of eIF4G1 $_{1-250}$ construct we observed that the second nickel column affects negatively to the protein stability and aggregation, therefore we purify the protein away from uncleaved, txA and TEV using a cation exchange column (SP 5ml). Finally proteins were concentrated and buffer exchanged according to their posterior use.

SEC-MALS measurements

The SEC measurements were carried out at room temperature, using a chromatographic system consisting of a LC-20AD pump, a RID-20A Differential Refractive Index Detector and a Dawn Heleos II MALS detector equipped with a K5 cell and laser (λ = 658 nm). For chromatographic separations, a Superdex 200 10/300 GL column was used at an eluent flow rate of 0.5 ml/min. For Pub1 RMM123 sample preparation, 6 mg in 25 mM KPi pH 6.5, 150 mM NaCl and 0.25 mM DTT buffer was filtered through 0.1 μm filter. In case of Pab1 RRM12 were used 12.3 mg in the same buffer and conditions. The proteins were centrifugated 15' at 10000 rpm at 4 ºC, then a total of 70 μl of the samples filtered were injected. The data were collected and analyzed by using the ASTRA SEC-software.

NMR experiments

All samples were prepared in NMR buffer (25 mM Phosphate pH 6.5, 25 mM NaCl, 1mM DTT and 10 % D_2O) and experiments acquired at 25 ºC on cryoprobe-equipped Bruker AV800 MHz spectrometer. The assignment of the backbone ^1H, ^{15}N and ^{13}C atoms was achieved by following the standard methodology. The 3D experiments 3D HNCO, 3D HNCA, 3D HC(CO)CA [29], 3D HN(CA)CO [30], 3D CBCANH [31], 3D CBCA(CO)NH [32] and 3D (H)CCH-TOCSY which were recorded to assign the side chain resonances [33]. Proteins concentration range between 100 – 200 μM. The chemical shifts of eIF4G1 $_{1-250}$, Pab1 RRM12 and Pub1 RRM123 will be deposited in the Biomagnetic Resonance Database (BMRB). Measurements of ^{15}N backbone amide relaxation T_1 and T_2 were measured with standard inversion-recovery and CPMG spin echo series of ^1H-^{15}N spectra. NMR spectra were processed using TOPSPIN v2.1 (Bruker, Inc) and NMRPipe [34], and they were analyzed with the program CcpNmr Analysis v2.4.2 [35].

Nitroxyl spin labelling and Paramagnetic Relaxation Enhancement (PRE) measurements

Protein sample from cysteine eIF4G $_{1-250}$ mutant (Q^{109}C) was chemically modified with the following protocol. Between 600 – 700 µM of the sample was exhaustively reduced with 5 mM DTT for two hours at room temperature. The DTT was eliminated by fast buffer exchange to 25 mM Tris pH 9.0 and 25 mM NaCl using desalting column.

Labeling with 4-(2-Iodoacetamido)-TEMPO was initiated immediately after column elution by adding a tenfold molar excess of the spin label dissolved in ethanol (25 mM spin label stock). Reaction was lead to proceed for 30 minutes at room temperature at darkness. The excess of 4-(2-Iodoacetamido)-TEMPO was quenched with 10 mM of 2-mercaptoethanol for 10 minutes and afterwards the protein adduct was exchange to 25 mM Phosphate pH 6.5, 25 mM NaCl and 1 mM DTT for later use.

The NMR samples were prepared in 5 mm sealed tubes in a N$_2$ atmosphere to avoid oxidation by air and high resolution 2D ^{1}H-^{15}N HSQC were recorded for the reduced state (active spin label). After that, the spin label was oxidized with 10 µM ascorbate [36] to record the reference 2D ^{1}H-^{15}N HSQC without paramagnetic relaxation enhancement. The relaxation effect was calculated as the intensity rations between peaks in the two spectra.

6.6. ACKNOWLEDGMENTS

6.7.   REFERENCES

1.   Sachs AB, Bond MW, Kornberg RD. A Single Gene From Yeast for Both Nuclear and Cytoplasmic Polyadenylate-Binding Proteins : Domain Structure and Expression. 1986;45:827–35.

2.   Adam SA, Nakagawa T, Swanson MS, Woodruff TK, Dreyfuss G. mRNA Polyadenylate-Binding Protein : Gene Isolation and Sequencing and Identification of a Ribonucleoprotein Consensus Sequence. 1986;6(8):2932–43.

3.   Ho B, Baryshnikova A, Brown GW. Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. Cell Syst. 2018;6(2):192-205.e3. Available from: http://www. sciencedirect.com/science/article/pii/S240547121730546X

4.   Anderson P, Kedersha N. RNA granules : post-transcriptional and epigenetic modulators of gene expression. Nat Publ Gr. 2009;10(6):430–6. Available from: http://dx.doi.org/10.1038/nrm2694

5.   Santiveri CM, Mirassou Y, Rico-Lastres P, Martinez-Lumbreras S, Perez-Canadillas JM. Pub1p C-terminal RRM domain interacts with Tif4631p through a conserved region neighbouring the Pab1p binding site. PLoS One. 2011/09/21. 2011;6(9):e24481. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21931728

6.   Brambilla M, Branduardi P, Martani F, Bertacchi S. The Saccharomyces cerevisiae poly (A) binding protein (Pab1): Master regulator of mRNA metabolism and cell physiology. 2019;(December 2017):23–34.

7.   Safaee N, Kozlov G, Noronha AM, Xie J, Wilds CJ, Gehring K. Interdomain allostery promotes assembly of the poly(A) mRNA complex with PABP and eIF4G. Mol Cell. 2012/10/09. 2012;48(3):375–86. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23041282

8.   Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res . 2018/05/23. 2018;46(W1):W296–303. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29788355

9.   Schäfer IB, Yamashita M, Schuller JM, Schüssler S, Reichelt P, Strauss M, et al. Molecular Basis for poly(A) RNP Architecture and Recognition by the Pan2-Pan3 Deadenylase. Cell. 2019;177(6):1619-1631.e21.

10.   Wallace EWJ, Kear-scott JL, Pilipenko E V, Schwartz MH, Laskowski PR, Rojek AE, et al. Reversible, specific, active aggregates of endogenous proteins assemble upon heat stress. Cell. 2015;162(6):1286–98.

11.   Buchan JR, Yoon JH, Parker R. Stress-specific composition, assembly and kinetics of stress granules in Saccharomyces cerevisiae. J Cell Sci. 2010/12/22. 2011;124(Pt 2):228–39. Available from: https://www. ncbi.nlm.nih.gov/pubmed/21172806

12.   Riback JA, Katanski CD, Kear-Scott JL, Pilipenko E V, Rojek AE, Sosnick TR, et al. Stress-Triggered Phase Separation Is an Adaptive, Evolutionarily Tuned Response. Cell. 2017/03/12. 2017;168(6):1028-1040 e19. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28283059

13.   Pérez I, McAfee JG, Patton JG. Multiple RRMs contribute to RNA binding specificity and affinity for polypyrimidine tract binding protein. Biochemistry. 1997;36(39):11881–90.

14. Duttagupta R, Tian B, Wilusz CJ, Khounh DT, Soteropoulos P, Ouyang M, et al. Global analysis of Pub1p targets reveals a coordinate control of gene expression through modulation of binding and stability. Mol Cell Biol. 2005 Jul;25(13):5499–513. Available from: https://www.ncbi.nlm.nih.gov/pubmed/15964806

15. Li H, Shi H, Wang H, Zhu Z, Li X, Gao Y, et al. Crystal structure of the two N-terminal RRM domains of Pub1 and the poly (U) -binding properties of Pub1. 2010;171:291–7.

16. Anderson JT, Paddy MR, Swanson MS. PUB1 Is a Major Nuclear and Cytoplasmic Polyadenylated RNA-Binding Protein in Saccharomyces cerevisiae. 1993;13(10):6102–13.

17. Buchan JR, Muhlrad D, Parker R. P bodies promote stress granule assembly in Saccharomyces cerevisiae. J Cell Biol. 2008/11/05. 2008;183(3):441–55. Available from: https://www.ncbi.nlm.nih.gov/pubmed/18981231

18. Kroschwald S, Munder MC, Maharana S, Franzmann TM, Richter D, Ruer M, et al. Different Material States of Pub1 Condensates Define Distinct Modes of Stress Adaptation and Recovery. Cell Rep. 2018/06/14. 2018;23(11):3327–39. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29898402

19. Martinez-Lumbreras S. The role of Gbp2p , Nab2p and Pub1p along the mRNA cycle : structural and molecular recognition studies by NMR and other biophysical techniques (Doctoral thesis). Univ Autónoma Madrid. 2013;

20. Yao G, Chiang YC, Zhang C, Lee DJ, Laue TM, Denis CL. PAB1 self-association precludes its binding to poly(A), thereby accelerating CCR4 deadenylation in vivo. Mol Cell Biol. 2007/07/11. 2007;27(17):6243–53. Available from: https://www.ncbi.nlm.nih.gov/pubmed/17620415

21. Brune C, Munchel SE, Fischer N, Podtelejnikov A V., Weis K. Yeast poly(A)-binding protein Pab1 shuttles between the nucleus and the cytoplasm and functions in mRNA export. Rna. 2005;11(4):517–31.

22. Handa N, Kukimoto-niino M, Akasaka R, Kishishita S, Murayama K, Terada T, et al. The Crystal Structure of Mouse Nup35 Reveals Atypical RNP Motifs and Novel Homodimerization of the RRM Domain. 2006;114–24.

23. Ge H, Zhou D, Tong S, Gao Y, Teng M, Niu L. Crystal structure and possible dimerization of the single RRM of human PABPN1. Proteins Struct Funct Genet. 2008;71(3):1539–45.

24. Sagnol S, Yang Y, Bessin Y, Allemand F, Hapkova I, Notarnicola C, et al. Homodimerization of RBPMS2 through a new RRM-interaction motif is necessary to control smooth muscle plasticity. Nucleic Acids Res. 2014/07/26. 2014 Sep;42(15):10173–84. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25064856

25. Teplova M, Farazi TA, Tuschl T, Patel DJ. Structural basis underlying CAC RNA recognition by the RRM domain of dimeric RNA-binding protein RBPMS. Q Rev Biophys. 2015/09/08. 2016 Jan;49:e1–e1. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26347403

26. Ripin N, Boudet J, Duszczyk MM, Hinniger A, Faller M, Krepl M, et al. Molecular basis for AU-rich element recognition and dimerization by the HuR C-terminal RRM. Proc Natl Acad Sci U S A. 2019/02/04. 2019 Feb 19;116(8):2935–44. Available from: https://www.ncbi.nlm.nih.gov/pubmed/30718402

27.   Neidhardt FC, Bloch PL, Smith DF. Culture medium for enterobacteria. J Bacteriol. 1974;119(3):736–47.

28.   Lin MT, Fukazawa R, Miyajima-nakano Y, Matsushita S, Choi SK, Iwasaki T, et al. Escherichia coli Auxotroph Host Strains for Amino Acid-Selective Isotope Labeling of Recombinant Proteins. 1st ed. Vol. 565, Isotope labeling of biomolecules labeling Methods. Elsevier Inc.; 2015. 45–66 p. Available from: http://dx.doi.org /10.1016/bs.mie.2015.05.012

29.   Grzesiek S, Bax A. Improved 3D triple-resonance NMR techniques applied to a 31 kDa protein. J Magn Reson. 1992;96(2):432–40.

30.   Clubb RT, Thanabal V, Wagner G. A constant-time three-dimensional triple-resonance pulse scheme to correlate intraresidue 1HN, 15N, and 13C' chemical shifts in 15N13C-labelled proteins. J Magn Reson. 1992;97(1):213–7.

31.   Grzesiek S, Bax A. An efficient experiment for sequential backbone assignment of medium-sized isotopically enriched proteins. J Magn Reson. 1992;99(1):201–7.

32.   Grzesiekt S, Bax A. Correlating Backbone Amide and Side Chain Resonances in Larger Proteins by Multiple Relayed Triple Resonance NMR. J Am Chem Soc. 1992;114(16):6291–3.

33.   Sattler M, Schleucher J, Griesinger C. Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. Prog Nucl Magn Reson Spectrosc. 1999;34(2):93–158.

34.   Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. NMRPipe: A multidimensional spectral processing system based on UNIX pipes. J Biomol NMR. 1995;6(3):277–93. Available from: http://www.scopus.com/ inward/ record.url?eid=2s2.00029400480&partnerID=40&md5=2ca830f73de65c42691425cd22747e94

35.   Skinner SP, Fogh RH, Boucher W, Ragan TJ, Mureddu LG, Vuister GW. CcpNmr AnalysisAssign : a flexible platform for integrated NMR analysis. J Biomol NMR. 2016;66(2):111–24.

36.   Gillespie JR, Shortle D. Characterization of long-range structure in the denatured state of staphylococcal nuclease. I. Paramagnetic relaxation enhancement by nitroxide spin labels. J Mol Biol. 1997;268(1):158–69.

6.8. SUPPORTING INFORMATION
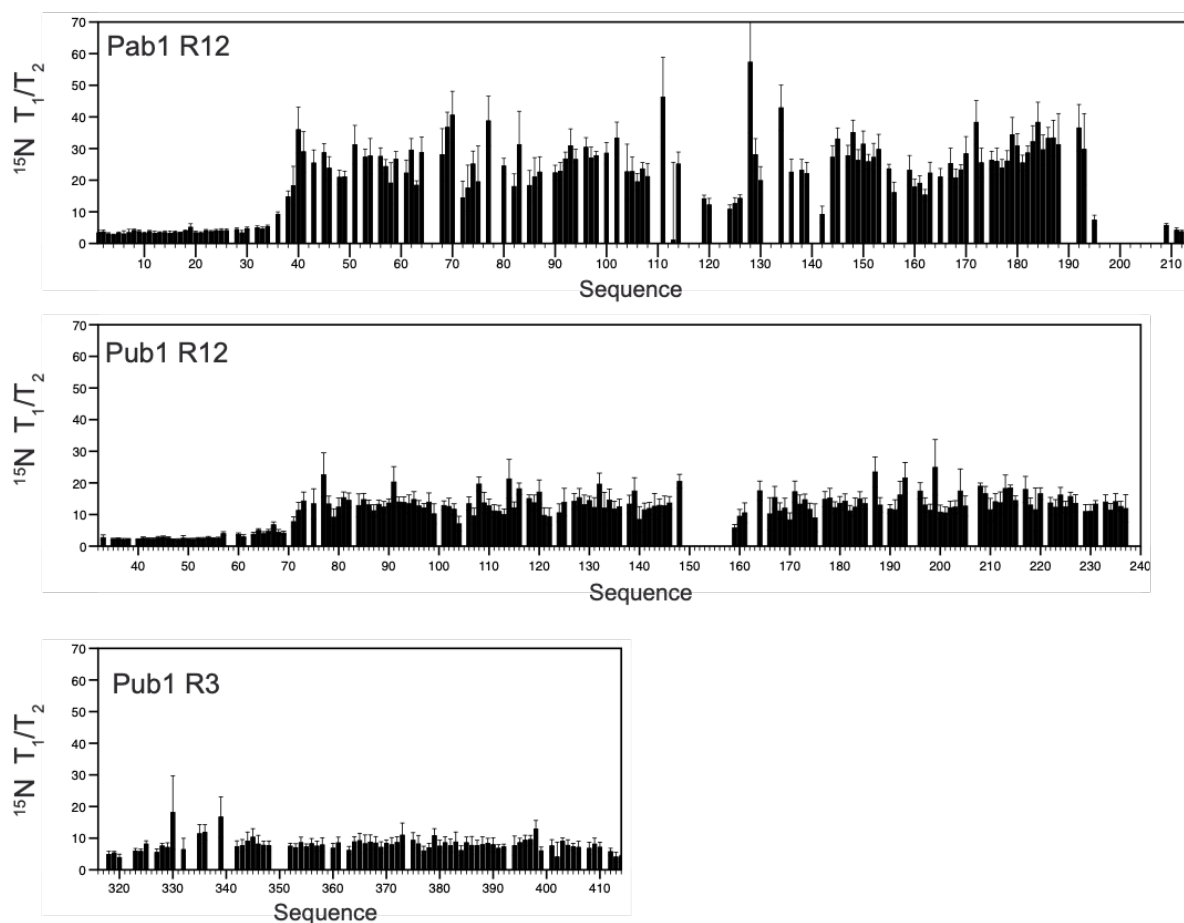
Figure S1. Dynamical analysis of Pab1 RRM1-RRM2 (Pab1 R12), Pub1 RRM1-RRM2 (Pub1 R12) and Pub1 RRM3 (Pub1 R3) thought ^{15}N relaxation data (T_1/T_2 ratio) analysis.
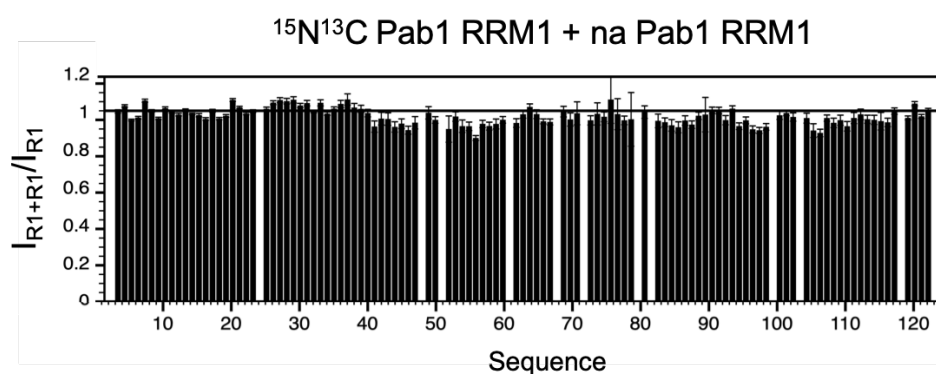
Figure S2. Intensity ratios of ^{15}N^{13}C Pab1 RRM1 and natural abundance (na) Pab1 RRM1 at 100 μM and the same buffer conditions. IDP part of Pab1 correspond with 1 – 40 approximately.
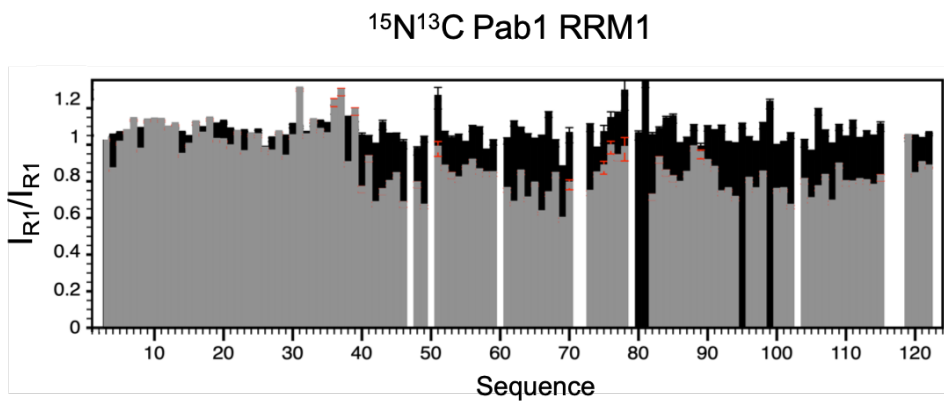
Figure S3. Intensity ratios of $^{15}N^{13}C$ Pab1 RRM1 at different concentrations. Pab1 RRM1 300μM and Pab1 RRM1 100 μM ratio (black histogram), and Pab1 RRM1 1000μM and Pab1 RRM1 100 μM ratio (grey histogram). IDP part of Pab1 correspond with 1 – 40 approximately.
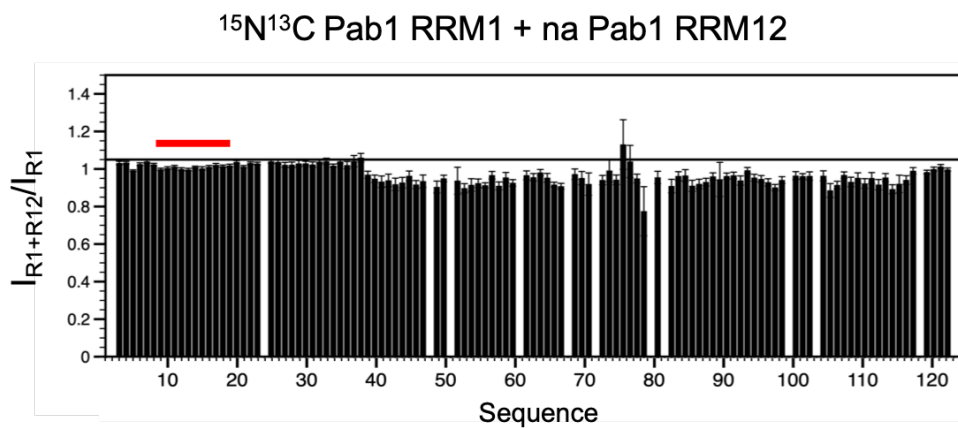
Figure S4. Intensity ratios of $^{15}N^{13}C$ Pab1 RRM1 and natural abundance (na) Pab1 RRM12 at 100 μM and the same buffer conditions. The red bar represents the transient contact between the IDP and RRM2 of Pab1.

CHAPTER 7

Multivalent interactions between eIF4G1 N-terminal IDP, Pab1 and Pub1 guide the formation of high order molecular species and protein condensates

7.1. ABSTRACT

The multivalent interactions between proteins play a key role in gene expression regulation processes such as translation, in which a high amount of proteins are involved. A crucial element in the translation process is eIF4G1, that acts as a scaffold by interacting with many other proteins . This study provides relevant structural information using NMR spectroscopy about some protein-protein interactions: (i) Pab1 and eIF4G1 $_{1-250}$ in which is recognized a new binding site that could have an important role in the cap-independent translation initiation, (ii) Pub1 and eIF4G1 $_{1-250}$ in which it has been determined the YNN recognition motif through the structure of a eIF4G1 $_{35-51}$ – Pub1 RRM3 chimera.

The eIF4G1-Pab1 interaction promotes eIF4G1 oligomerization detectable as condensates using confocal microscopy. These high order structures recruit Pub1 protein to become part of these condensates. The binding sites of Pab1 and Pub1 with eIF4G1 $_{1-250}$ do not overlap, and both proteins do not interact. The simultaneous binding of these three proteins seems to change the conformational landscape of eIF4G1 $_{1-250}$ enhancing it's the prion-like propensities which may result in the formation of these condensates. Moreover, a possible new RNA interaction motif (QQQRXΦ) has been identified to interact on the RNA binding domain of (RNA1) of eIF4G1 $_{1-250}$. Finally, all the information has been put together in a structural model that aims to unravel some of the structural details of the SG cores.

Keywords: Stress granules (SG), multivalent interactions, Confocal microscopy, Nuclear Magnetic Resonance (NMR), cap-independent translation

7.2. INTRODUCTION

Translation is essential for gene expression that is heavily regulated by different mechanisms and signalling networks involve dozens of proteins [1–4]. One of these proteins is eIF4G1 that is a central player in the cap-dependent mechanism of transcription initiation [5,6] and paradoxically also in the composition of stress granules [7–9]. This ambivalent nature, capable to activate or repress translation depending on the circumstances, is a common trait of other translation factors and RNA binding proteins (RBPs) that participate in these processes. Many of these proteins contain intrinsically disordered regions that can undergo post-translational chemical modifications (e.g. phosphorylation and methylation) and physical transitions (i.e. liquid-liquid phase separations LLPS), which are often linked on each other [10]. In this sense, these type of proteins behave as sensors capable to evolve to condensed phases upon different stressors, that are likely at the origin of stress granules and translation arrest [11].

Pab1, Pub1 are eIF4G-binding proteins and can behaves as "stress-sensors" triggering the LLPS response. At the same time, they play important roles in translation initiation. This duality raises important questions about how molecular interactions involving proteins and eIF4G1 regulate translation initiation (i) and nucleation of stress granules (ii) and which are the key elements tip the balance in favour of translation activation or repression.

eIF4G1 interacts with Pab1 [5] and eIF4E [12] to form the "closed-loop" structure that is the most widely accepted mechanism of functional communication between $5' - to - 3'$ ends of the mRNAs [12]. However, no all mRNAs require this structure for translation and 5'-3' circularization can be achieved by other means [13]. The interaction between Pab1 and eIF4G1 is conserved from yeast to humans and structure of the human ternary complex eIF4G:PABP:A[12] enlighten the chemical basis of cap-dependent translation [14]. In other cases, mRNAs contain internal ribosome entry sites (IRES) on their 5'-UTR, which have poly(A) tracts that can bind to Pab1 [15]. On these cases mRNA circularization can be mediated by dimerization between Pab1 molecules bound to 5'-UTR and 3'-poly(A) tail. This and other cap-independent mechanisms require eIF4G but not the cap-recognition activity of eIF4E.

In contrast to Pab1, Pub1 seems to have an antagonist role in translation. Pub1 is present in preinitiation complexes and in monosomes but not stable associated with polysomes [16]. This suggests that it is removed during the first round of translation, or shortly after translation initiation. Pub1 is an abundant protein that binds to ~6% of the yeast transcripts being important for their stability [17]. Many of these transcripts involved in stress response. Indeed, the principal role of Pub1 is related to the stress response and the formation of stress granules (SG), micrometer – size membrane – less sub-organelles that highjack the components of the translation machineries in an inactivate state. Like Pab1 and eIF4GI, Pub1 is bona-fide SG marker and has a pH/temperature induced LLPS response that postulates it a stress-biosensor [11,18]. Pab1 has been reported to have similar LLPS properties [19].

The formation and disaggregation of SG is in part regulated by signaling pathways [20], but their onset is heavily influenced by the physicochemical properties of some of the protein components of the cytoplasmic bodies. *In vitro* biophysical studies are necessary to understand which are the factors that govern the LLPS behavior, like recently for the cases of Pub1[11] and Pab1[17]. But the next and more challenging step is to perform similar studies with more complex mixtures, in order to understand the interplay between SG components (proteins and RNAs). More importantly it is necessary to reveal the process a higher resolution, using techniques as NMR that allow to get information at residue/atomic level.

This chapter studies by NMR the interaction between Pab1 and Pub1 constructs with eIF4G1$_{1-250}$. The results reveal a complex landscape of protein binding sites and suggest the formation of high order structures upon different combinations. The experiments are complemented with turbidity and confocal fluorescence microscopy studies that allowed to follow up the formation of micrometer-size particles. The results suggest that multivalence interactions between Pub1/Pab1/eIF4G are sufficient to assemble protein – protein condensates even in the absence of pH/temperature stresses. In combination with the results of chapter 5 and 6, a model that progress in our understanding about the complex biophysics behind the assembly of membrane-less organelles is proposed.

7.3. RESULTS AND DISCUSSION

Structural characterization of the Pub1/eIF4G interaction

The interaction between eIF4G1 $_{1-187}$ and Pub1 RRM3 was previously characterized by NMR [21,22]. Pub1 RRM3 interacts with RNA1_1 and BOX1 regions as shown by using peptide probes. Here we monitor the Pub1 RRM3-eIF4G1 $_{1-250}$ interaction on the eIF4G1 $_{1-250}$ spectra construct and found that the perturbed residues remain at the N-terminus (**Figure 1A**). This indicates that BOX3, the putative Pab1 binding site, does not participate in Pub1 recognition, and its long-range contacts seems to remain unchanged. On the other hand, Pub1 RRM3 binding to RNA1_1 and BOX1 seems to affect to other regions like RNA1_2, likely through subtle rearrangements of the conformational landscape, as previously described [21,22].

The binding of BOX1 and RNA1_1 peptides is too weak and prevents the detection of intermolecular NOEs in the 2D NOESY, making unfeasible the calculation of the NMR structure of the complex. To overcome this technical problem, we produced recombinant chimeras with the RNA1_1 sequence fused to N- or C-terminus of Pub1 RRM3. In this way the interaction between the two parts becomes a folding event that is concentration independent and energetically more efficient. The ^{1}H-^{15}N HSQC spectra of the configuration with the peptide fused to the C-terminus of Pub1 (RRM3 domain) is similar to that of Pub1 RRM3 alone, whereas the N-terminal chimera differs significantly, suggesting that only in this case the peptide can effectively fold-back into the binding site (**Figure 1B**), while in the first case the interaction is probably not geometrically possible.

Data was collected to calculate the structure of the eIF4G1 $_{35-51}$ – Pub1 RRM3 chimera, which enlighten the key elements for recognition (**Figure 1C**). The structure reveals that eIF4G1 residues Y^{41}, N^{42} and N^{43} (part of the conserved motif mentioned before) interact with a swallow cleft defined by the contact between helix α1 and strand β2 of the Pub1 RRM3. The tyrosine is inserted in a small cavity and the two asparagine's point outwards. The interaction surface is small (buried ASA ~400 Å^2) , which explains why the eIF4G1 $_{35-51}$ – Pub1 interaction is weak.

Figure 1. **(A)** Intensity ratios of ^{15}N eIF4G1 $_{1-250}$ (eIF4G1) and natural abundance Pub1 RRM3 (Pub1 R3). The schematic representation of the conserved sequence boxes of eIF4G1 $_{1-250}$ are represented on the top. **(B)** Superpositions of 2D ^{1}H-^{15}N HSQC of Pub1 RRM3 free (black) and the spectra of chimeras eIF4G1$_{35-51}$-Pub1 RRM3 (red) and Pub1 RRM3-eIF4G1$_{35-51}$ (red). **(C)** NMR structure of the eIF4G1$_{35-51}$-Pub1 RRM3 chimera with the eIF4G1 part in yellow and Pub1 RRM3 in white/blue. Key interacting residues have been labelled on both regions.

eIF4G1 $_{1\text{-}250}$ contains two Pab1 binding sites

Next, we studied the interaction between eIF4G1$_{1\text{-}250}$ and Pab1. We titrated unlabelled Pab1 RRM12 over ^{15}N labelled eIF4G1$_{1\text{-}250}$ (**Figure 2A**) causing perturbations and signal disappearance in three well-delimited regions: 95 − 103 (BOX1), 135 − 160 (BOX2) and 200 − 234 (BOX3). A similar signal disappearing pattern is observed when following the titration with isotopically discriminated NMR spectroscopy [23,24] on different amino acid-selective unlabelling samples (**Figure S2**). Until this work only the BOX3 was reported as Pab1 binding site [25,26], therefore the changes on the other two boxes were unanticipated. Titration with peptides corresponding to these regions on ^{15}N- labelled Pab1 RRM12 shows that only BOX3 and BOX2 cause changes on Pab1 spectra, that are more extensive with BOX3 (the canonical site) (**Figure 2B,C and Figure S3**). This suggest that the changes observed on BOX1 for the eIF4G1$_{1\text{-}250}$ are due to indirect rearrangements of internal contacts. As predicted by previous studies [4,27], only Pab1 RRM2 is involved in recognition of both BOX2 and BOX3, and the interaction site in yeast coincides with that in human PABP1 [28]. The peptide − BOX2 causes less changes and probably interact weaklier (**Figure S4**), that in the context of eIF4G1$_{1\text{-}250}$. Like for BOX3, several signals in BOX2 and surroundings, disappear upon titration. This suggest that Pab1 RRM2 forms high molecular weight complexes with eIF4G1. A proposed mechanism involves the cooperative recognition of BOX2 and BOX3 by Pab1 RRM12 dimers (**Figure 2D**).

The discovery of second Pab1 binding site in eIF4G1 IDD could have important implications for the role of Pab1 in translation initiation. The conservation of BOX2 was noticed before, but not biological role was assigned to it [29]. The cooperative recognition of the two Pab1 sites in eIF4G1 reinforces the "close-loop" model (**Figure 3A**), and at the same time allows to propose a mechanism for cap-independent translation initiation (**Figure 3B**). Some mRNAs that are activated upon nutrient deprivation contain A-rich sequences on their 5'-UTRs [30,31]. In contrast to the poly(A) tail, these 5'-UTR A-tracks are short and can only accommodate one Pab1. This could interact with the poly(A)-bound Pab1 by means of the double eIF4G1 motif and by Pab1 self-recognition, leading to 5'-3' mRNA circularization without the requirement of cap-binding. This model provides a mechanism for these selected genes when cap-dependent translation is repressed (e.g. binding of 4EBPs to eIF4E).
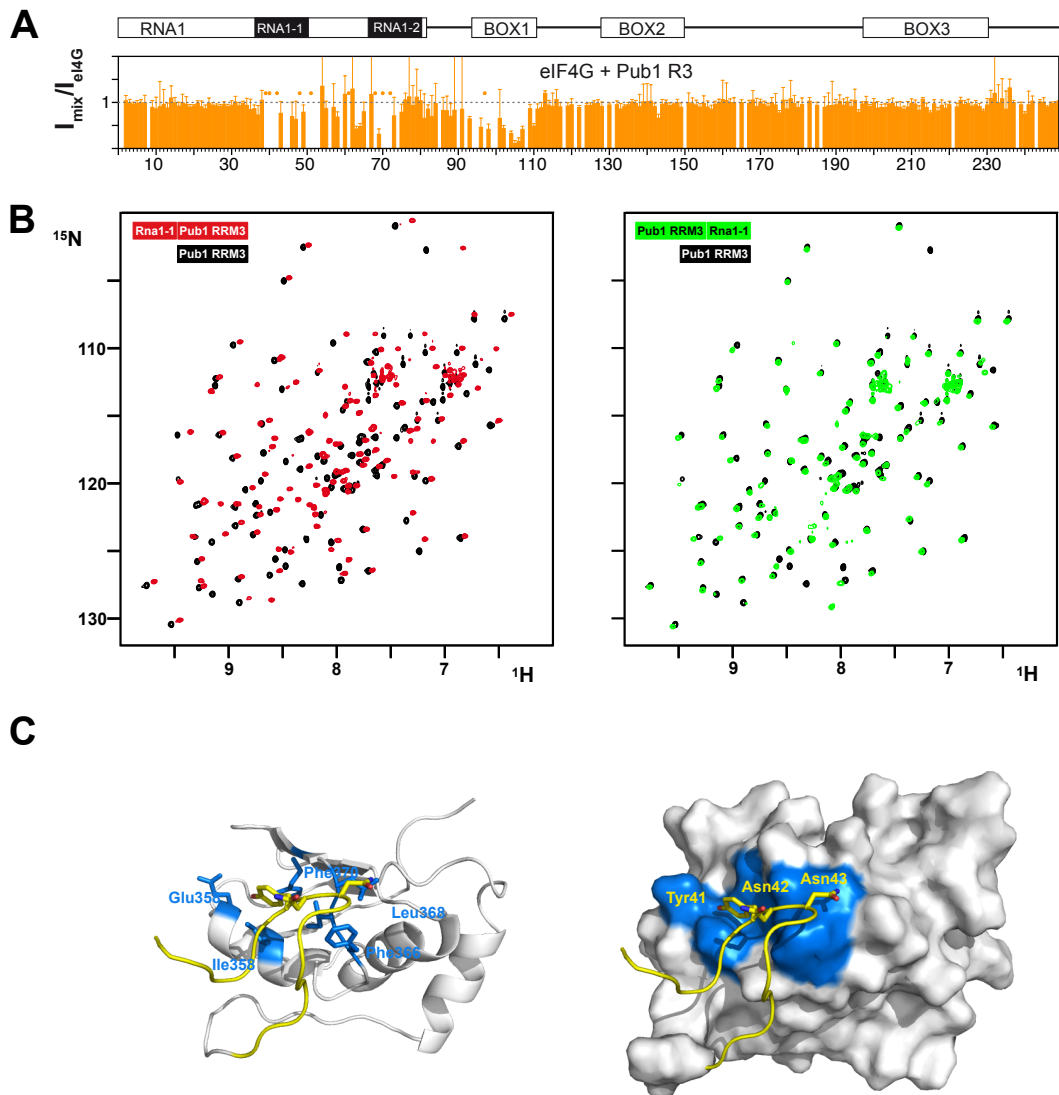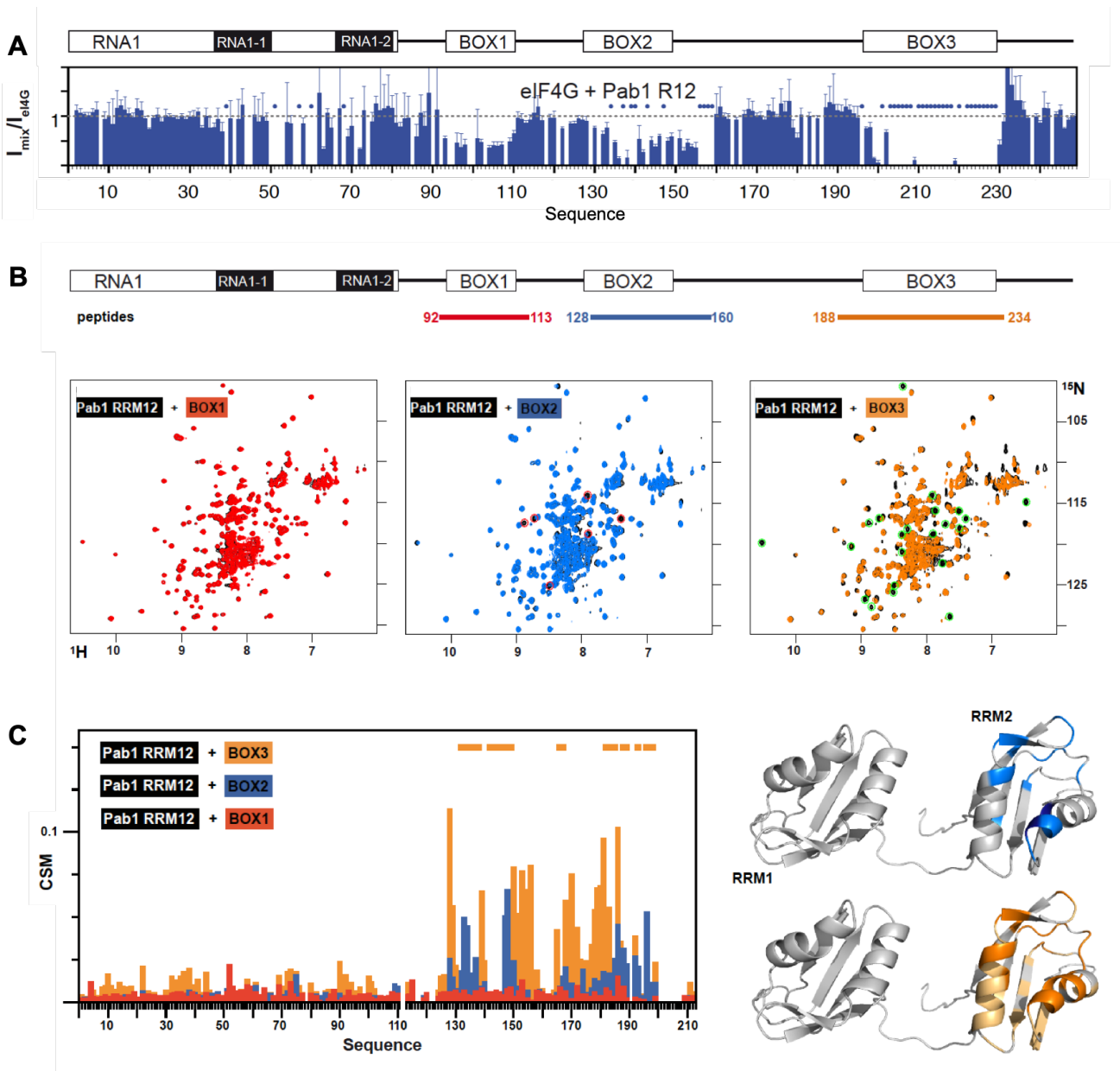
Figure 2. **(A)** Intensity ratios of [15]N eIF4G1 $_{1-250}$ (eIF4G1) and natural abundance Pab1 RRM12 (Pab1 R12). The schematic representation of the conserved sequence boxes of eIF4G1 $_{1-250}$ are represented on the top. **(B)** eIF4G1 $_{1-250}$ peptides (BOX1 in red, BOX2 in blue and BOX3 in orange) titration on 2D ^{1}H-^{15}N HSQC experiment of Pab1 RRM12 (black). **(C)** chemical shift mapping (CSM) of Pab1 RRM12 in its titration with eIF4G1 $_{1-250}$ peptides (BOX1 in red, BOX2 in blue and BOX3 in orange). These BOX1 and BOX2 mapping are showed on Pab1 RRM2 model (left model structures).
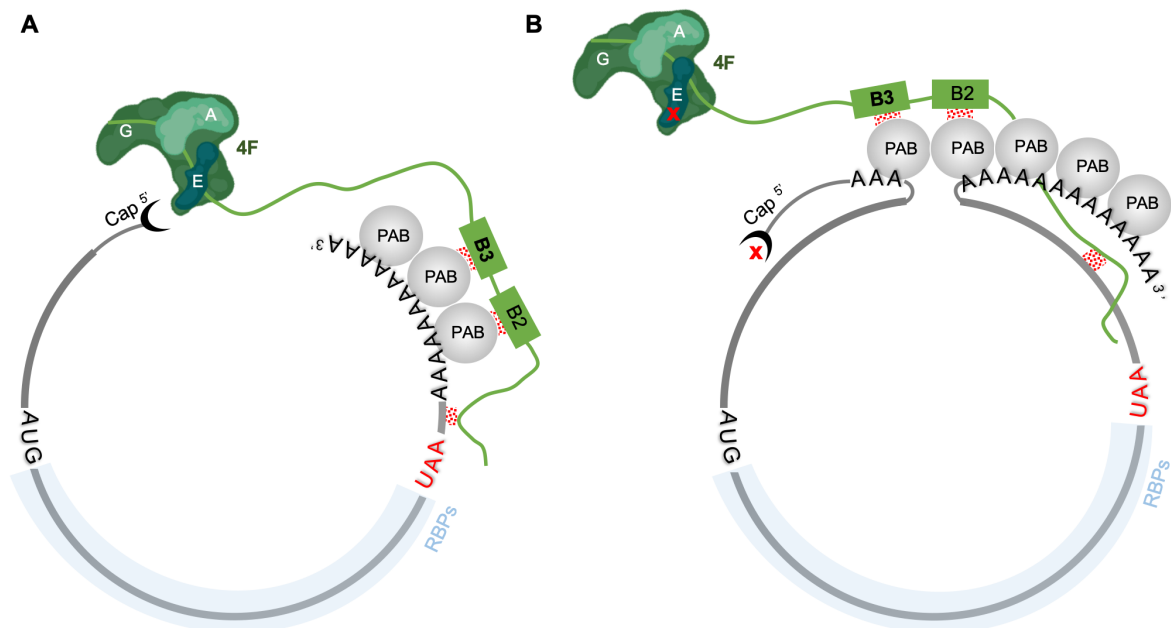
Figure 3. Different mechanism for translation initiation. (**A**) The close-loop model with the incorporation of the dual binding site for Pab1 in eIF4G. (**B**). A mechanism of cap-independent translation initiation in which circularization is promoted by Pab1 binding to 5'-UTR A-tracks.

Comparison of Pab1 and Pub1 recognition modes.

The crystal structure of human PABP with poly(A) and an eIF4G1 peptide represents an example of this key interaction for the close-loop mechanism [14]. The recognition by RRM takes place through an interface formed by the two α-helices of the domain (**Figure 4**). In contrast, the structure of the eIF4G1$_{35-51}$-Pub1 RRM3 chimera shows that Pub1 uses a different binding locus. This interface is smaller (~400 Å^2) than the PABP one (~1100 Å^2). Similar locations have been involved in RNA [32,33] and protein [34] recognition in other RRMs. Finally, the two proteins recognise eIF4G1 $_{1-250}$ through interfaces not overlapping with the RNA ones, suggesting that both events should be compatible.

The human eIF4G1 peptide that interacts with PABP is structurally disordered on its free state [14] and folds upon binding to form an α/β structure. In contrast, yeast eIF4G1 BOX3 forms an α-helix in the

free state, suggesting a different recognition mechanism. Moreover, the Pab1 binding site has four hydrophobic residues F_{209}, I_{210}, V_{213} and I_{215} that can play a similar role to those in the PABP/eIF4G/A_{10} structure (**Figure 4**), but the yeast peptide cannot be recognised by the human PABP [35], once again suggesting that binding modes might differ significantly between the two species.
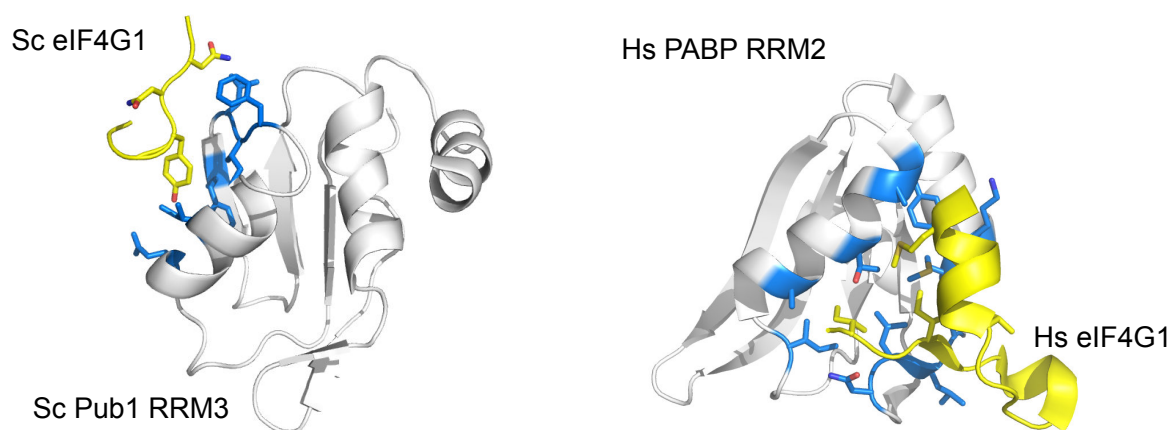
Figure 4: Comparison between eIF4G recognition modes of Pub1 RRM3 and Pab1 RRM2.

Simultaneous binding of Pub1 and Pab1 to eIF4G1 $_{1-250}$ trigger oligomerization

Previous studies in our lab found out that constructs eIF4G1 $_{1-82}$, eIF4G1 $_{1-305}$, eIF4G1 $_{1-348}$ and eIF4G1 $_{1-401}$ cause small changes on the Pub1 RRM3 spectra, whereas titration with eIF4G1 $_{1-190}$ causes the complete disappearance of all Pub1 RRM3 ^{1}H-^{15}N HSQC signals [22]. This effect disappeared when mutating the prion-like sequence in BOX1 [22]. In this thesis, the effect was analysed using the eIF4G1$_{1-250}$ construct. This binds to Pub1 RRM3 weakly (**Figure 5A**) and cause similar chemical shift perturbations that either other forms of eIF4G1 [20] or its peptides [27]. However, further titration with unlabelled Pab1 RRM12 reproduce Pub1 RRM3 signal disappearance (**Figure 5C**). Pab1 RRM12 binding sites do not overlap with Pub1 RRM3 ones and both proteins do not interact (**Figure 5B**). The simultaneous binding of both proteins seems to change the conformational landscape of eIF4G1 $_{1-250}$ enhancing its the prion-like propensities. The disappearance of Pub1 RRM3 signals suggest that it is more stable-associated with the oligomeric forms of eIF4G1$_{1-250}$.
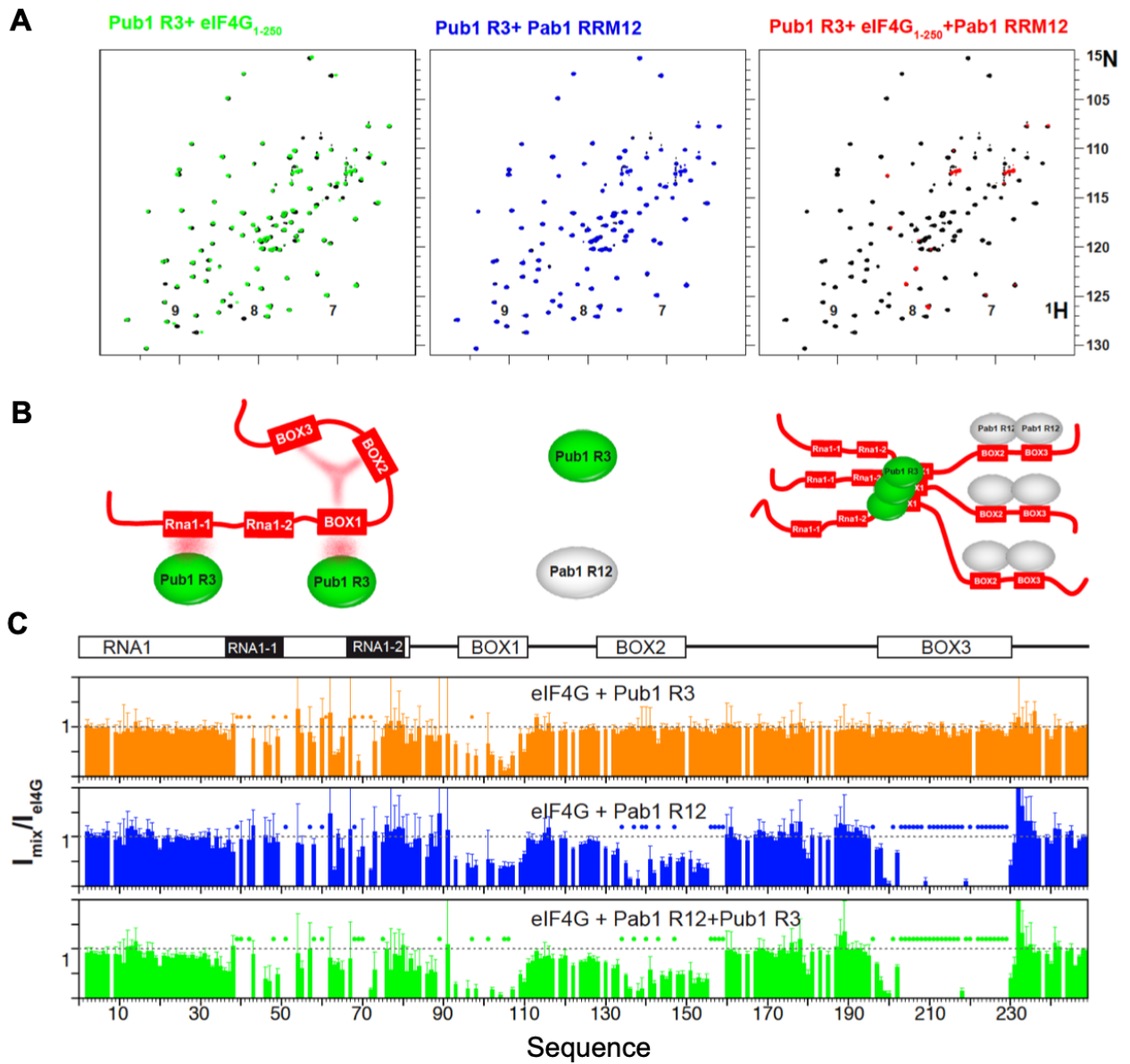
Figure 5. **(A)** Different binding events monitored on the Pub1 RRM3 spectra. Superposition of Pub1 RRM3 free state (black signals) with mixtures with eIF4G1$_{1-250}$ (on the left in green), Pab1 RRM12 (in the middle in red) and eIF4G1$_{1-250}$/Pab1 RRM12 (on the right in red). **(B)** Models depicting the different scenarios of the spectra shown above. (C) Different binding events monitored on the eIF4G1 $_{1-250}$ ^{1}H-^{15}N HSQC spectra. Intensity ratios between free eIF4G1 $_{1-250}$ and mixtures with Pub1 RRM3 (orange), Pab1 RRM12 (blue) and Pub1 RRM3/ Pab1 RRM12 (green). Dots over the sequence mark signals that disappear due to severe broadening.

The monitorization of the triple mixture on the eIF4G1$_{1\text{-}250}$ ^{1}H-^{15}N HSQC spectrum shows a combination of the effects of Pub1 RRM3 and Pab1 RRM2 binding, but with further intensity decrease of BOX1 signals, which is in agreements with the mechanism proposed (**Figure 5D**).

Altogether the data suggest that the combined effect of Pub1 RRM3 and Pab1 RRM12 promotes aggregation/oligomerization of eIF4G1. A possible explanation is that intramolecular interactions in eIF4G1$_{1\text{-}250}$ prevent BOX1 self-assembly in the free state and that the interactions with Pab1 and Pub1 remove these safeguards triggering eIF4G1 oligomerization.

The Pub1 RRM3 binding motif (YNN) and the prion-like motif (YYNN) of eIF4G1 overlap. Despite their simplicity, these two motifs are statistically underpopulated on the yeast proteome, but are relatively abundant among Pub1 and eIF4G1 binding proteins and on SG. Besides the two eIF4G isoforms, the YNN motif is found three RNA binding proteins (Nab6, Ngr1 and Npr1) (Figure S5). In the case of the prion-like sequence YYNN, it is found on Ksp1 (a serine kinase of the TORC pathway) and Slf1 (another RBP). Except for Nab6, all the target motifs map in predicted IDD, suggesting that these proteins might be recruited to the SG through similar interactions as described here for Pub1/eIF4G.

eIF4G1 $_{1\text{-}250}$ RNA recognition

Besides protein recognition, eIF4G1 binds RNA using three regions RNA1, RNA2 and RNA3 [29,36]. The construct eIF4G1 $_{1\text{-}250}$ contains one them (RNA1: residues 1-82). In this work, the NMR titrations on eIF4G1 $_{1\text{-}82}$ with different poly(A) oligos show that the strength of the interaction is length-dependent (**Figure 6 A**). Upon titrating with A^{12} a new signal appears in the spectra that corresponds to the signal of a side chain arginine guanidinium group (Nε-Hε) (**Figure 6 B**). This suggest that this chemical group is directly involved in RNA contacts, slowing down the otherwise rapid solvent exchange in the free protein (no Arg Nε-Hε signals). The interaction with poly(A) maps three regions centred around Arg$_{34}$, Arg$_{55}$ and Gly$_{65}$. Interestingly the first two, more strongly affected by the RNA, define a possible new RNA binding motif: QQQRXΦ (Φ=aromatic). The protein contains an RGG box (R$_{60}$GG), which is a well-known RNA binding sequence, but surprisingly it is less affected by binding than the novel sites. The observed length-dependence might reflect the possibility for the poly(A) to accommodate one or two binding sites.
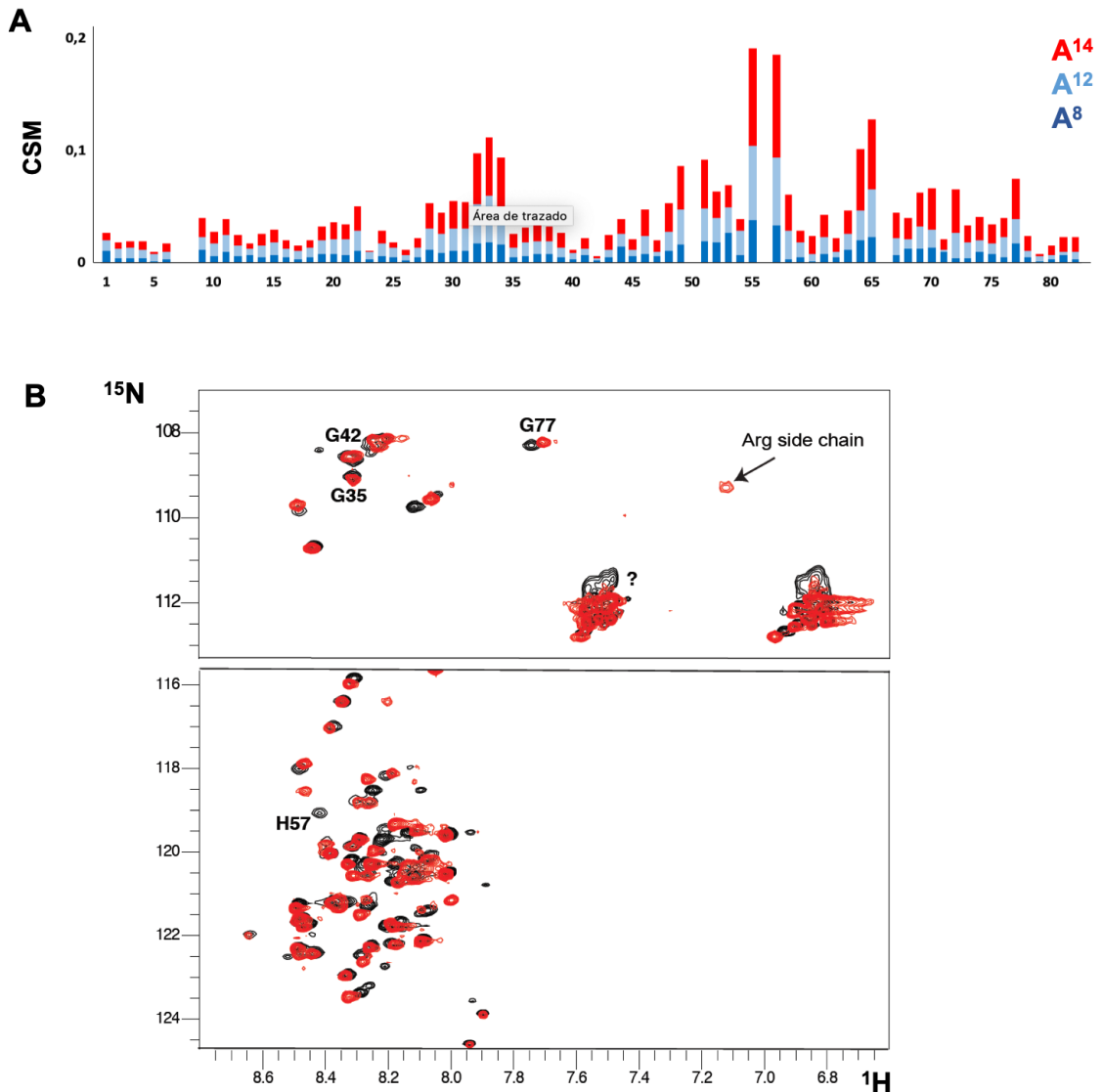
Figure 6: RNA recognition by eIF4G1$_{1-82}$. (A) Chemical shift mapping of various poly(A) probes. (B) Superposition of the ^{1}H-^{15}N HSQC of the free (black) and A$_{14}$-bound protein (red).

Pab1 – Pub1 – eIF4G1 $_{1-250}$ mixtures form micrometre-size condensates

Previous studies showed that Pub1, Pab1 and eIF4G1 proteins individually form liquid-liquid phase separations (LLPS) *in vitro* [11,19,37]. The formation of LLPS is thought to be promoted by disordered regions, however, it has also been suggested that the RRMs of Pab1 [19] and Pub1 [11] could also be involved. These studies have been performed on the individual proteins and have not explored the potential of heterotypic interactions between stress granules components.

Results in the previous section suggest that simultaneous binding of Pab1 RRM12, eIF4G1 $_{1-250}$ and Pub1 RRM3 lead to high order structures that become invisible to NMR because their size. Fluorescence microscopy was used to explore if Pub1/Pab1/eIF4G1 mixtures can actually form macroscopic condensates. The proteins were prepared at concentrations similar to *in vivo* and to previous experiments (see Materials and Methods) [11,19,37]. First, we used different crowders to simulate cell cytoplasm compactness in the three proteins sample (**Figure S6**). The confocal images of the triple mixture containing Alexa 488 labelled eIF4G1 (eIF4G1– Alexa 488) in 200 g/L Ficoll 70 revealed the presence of discrete round structures of small size (~1-2 µm, **Figure 7**), reminiscent of the condensates previously described for isolated Pab1 or Pub1 [11,19]. The size of these condensates appears to be smaller, but in this case, they are present without needing to apply pH or temperature shifts. Both, Pab1 and Pub1 colocalized with eIF4G1 $_{1-250}$ in these assemblies, as observed in images in which the proteins were pairwise labelled with spectrally different dyes (eIF4G1– Alexa 488/Pub1 – Alexa 647 or eIF4G1– Alexa 488/Pab1 – Alexa 647, **Figure 7**). Turbidity measurements performed in parallel with the unlabeled proteins in 200 g/L Ficoll show a modest (but reproducible) signal, compatible with the formation of relatively small structures (**Figure 7**). These structures were also formed in the presence of other crowding agents (Dextran 500 or PEG 8) as we mentioned before, while being scarce in their absence (i.e. in buffer) (**Figure S6**).

Similar structures were observed for binary mixtures of Pab1 and eIF4G1, in good agreement with the turbidity measured. In contrast, Pub1/eIF4G1 or Pub1/Pab1 mixtures show lower turbidity and the lack of visible structures in the confocal images **Figure 7 and S7**. Moreover, none of the three proteins formed on their own structures detectable by confocal microscopy or turbidity, in the presence of Ficoll (**Figure S7**). Our results show that the eIF4G1/Pab1/Pub1 mixtures can form crowding-driven structures resembling those previously described for full length Pab1 or Pub1 [11,19], but without pH or temperature stresses. Interestingly, Pub1 only becomes part of these condensates when simultaneously present with eIF4G1 and Pab1, a result that fully back up the NMR observations. Altogether the NMR and fluorescence microscopy data suggest that mixtures of Pub1, Pab1, eIF4G1 have an intrinsic propensity to form protein condensates. However, we tested if the condensates are formed conditions similar to the NMR experiments (proteins at ~100 µM and without crowders), but found no evidences of them. This suggest that the high other structures proposed in the NMR

experiments contain a discrete number of molecules and might represent an early stage of the condensate.
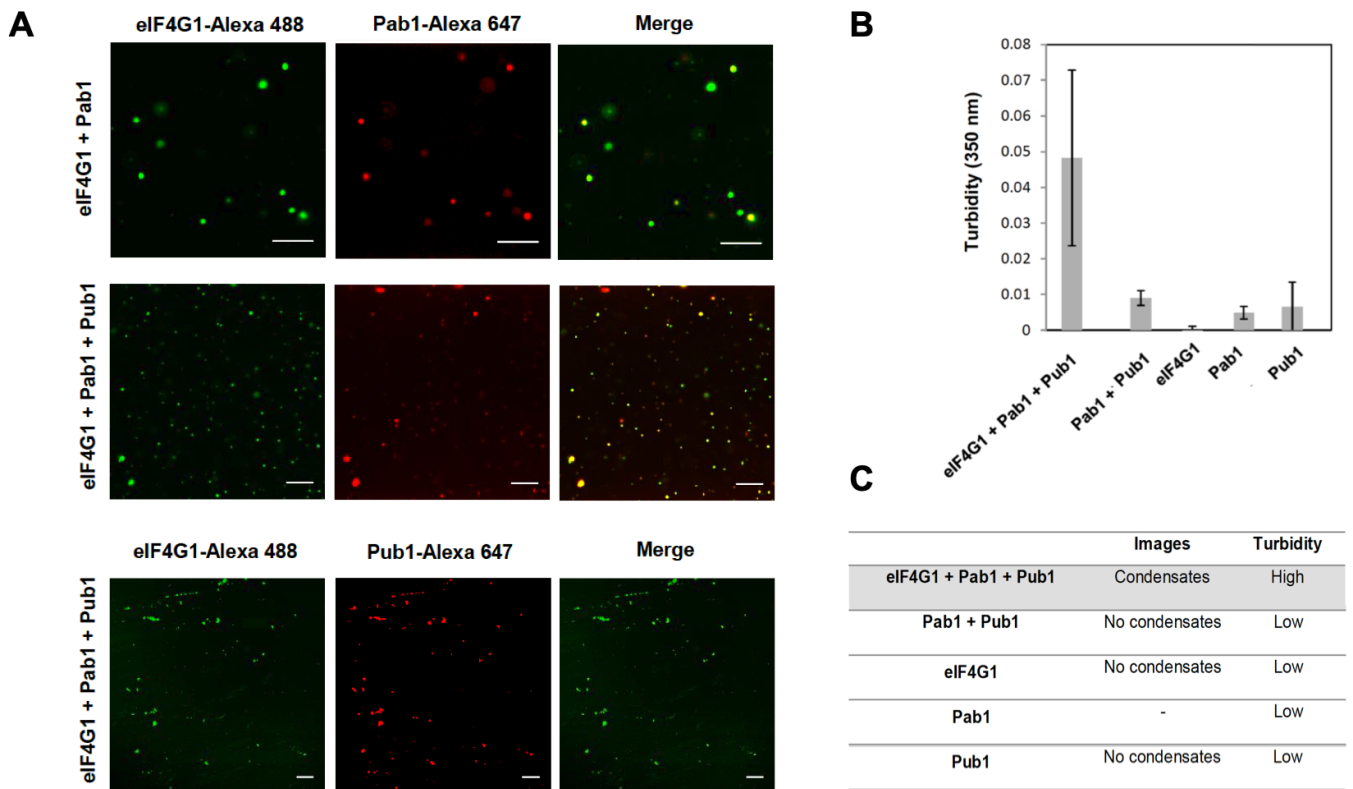
Figure 7. Some mixtures of Pab1, Pub1 and eIF4G1 form protein condensates *in vitro* (**A**) Confocal images of condensates attending at proteins used in 200 g/l Ficoll 70. (**B**) Turbidity measurements to monitor the formation of proteins condensates in PBS buffer and 200 g/l Ficoll 70 for different Pab1/Pub1/eIF4G mixtures. The protein concentrations as in (**A**) and data generated as averages of at least 3 independent measurements. (**C**) Table summarizing the A and B experiments.

A multivalent interaction model for biological condensates

Stress Granules and other biological condensates are complex structures that probably involved thousands of different classes of interactions. These cellular assembles are composed by many proteins, RNAs, pieces of machineries (i.e. ribosomes) and likely other small molecules and metabolites whose precise nature remain unknow. Moreover, they are dynamic structures, although it is becoming clear that they have a core structure that exchange components more slowly [38].

The structural and biophysical studies on these structures represent one of the challenges for the future and will need the concourse of multidisciplinary approaches form physics to biology, and from stated-of-the-art experiments to theoretical calculations. The situation now is focused in a broad definition of the types and nature of the interactions involved in the assemble of these condensates. It has been proposed that IDPs and RNA binding are important for LLPS but no precise mechanism have been derived at atomic level. Here we propose a number of multivalent interactions centered on eIF4G1 and its binding partners (Figure 8). These involve prion-like structures, protein-protein recognition involving folded domains, planar π-π and cation-π networks and RNA recognition by new motifs. All share in common that the take advantage of multivalence as a powerful mechanism to construct organized 3D networks. Because the interactions involved are weak, these networks can be very dynamic; simply by swapping the specific partners. And at the same time their vast number allow to form stable condensates.
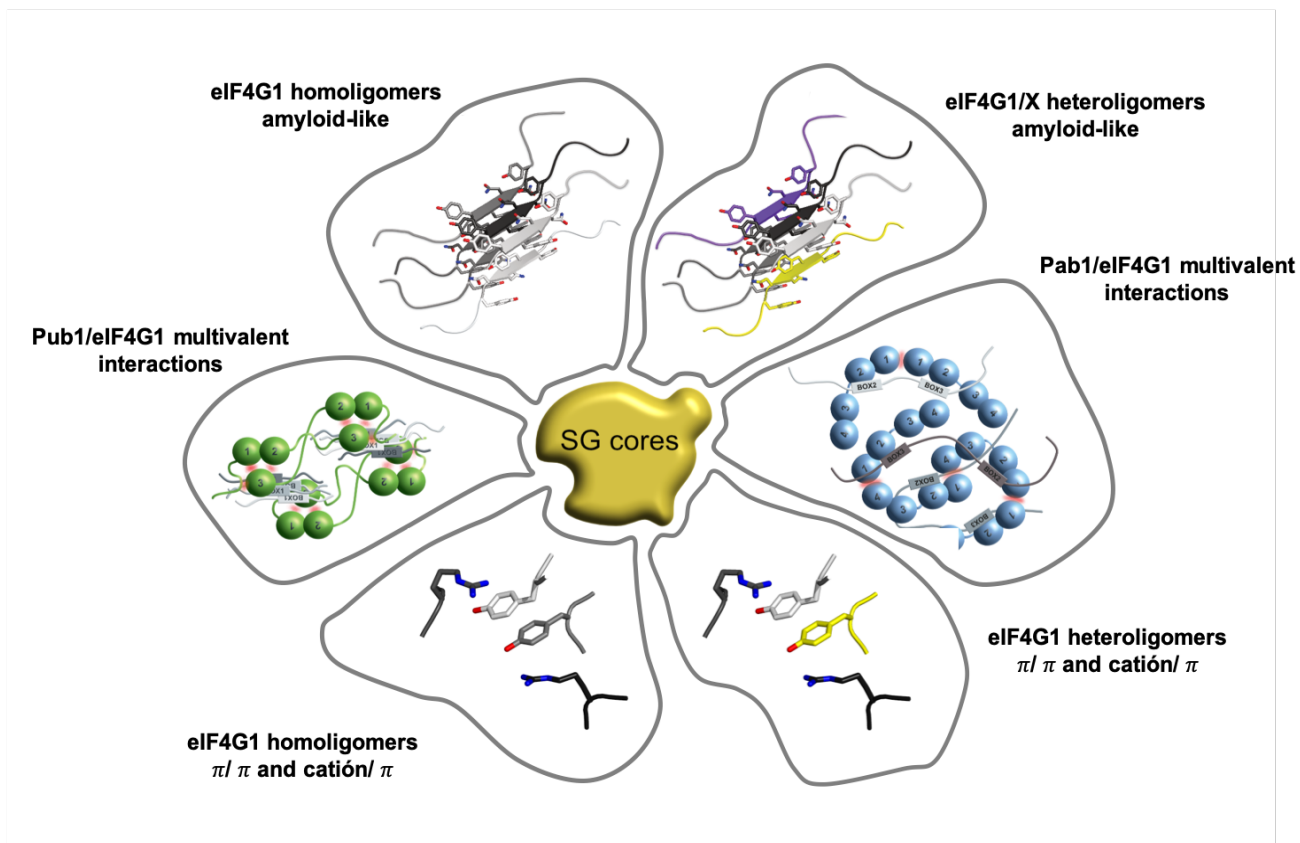
Figure 8: Multivalent interactions involving the eIF4G1 N-terminal IDD that are proposed to be important for the structure of SG cores.

7.4. CONCLUSIONS

Translation is a hugely complex process requiring the coordinated action of many factors, and consequently offers many targets for regulation [4]. Almost all eukaryotic mRNAs are thought to require the formation of a closed-loop complex for their efficient translation, where the cap-binding protein eIF4E and Pab1 recognize the 5' and 3' of mRNAs, and both proteins are connected by the scaffold protein eIF4G1 to start the translation [39]. However, there are mRNAs which can initiate translation through a cap – independent mechanism. In order to carry out this new process these mRNAs bind to Pab1 [15]. Here we describe that eIF4G1 $_{1-250}$ interacts with Pab1 in two sites (BOX2 and BOX3). The new binding site (BOX2) could suggest an important new Pab1 role in the cap – independent translation initiation. Moreover, as we mentioned in Introduction section, the principal role of the other protein of this study, Pub1, is to be an essential component in the formation of stress granules (SG). This chapter shows that the Pub1 – eIF4G1 $_{1-250}$ interaction is regulated by Pab1, although there is no direct interaction between Pub1 and Pab1. The binding of Pab1 to eIF4G is suggested to promote the formation of eIF4G1 oligomers (note that Pab1/eIF4G1 form condensates (**Figure 7**) to which Pub1 is readily recruited. This switching mechanism could be one of the ways for SG formation in which the unusual Pub1 RRM3 domain [21] could play a leading role.

Compared to previous studies, this one emphasizes the importance of multivalent interactions between different components of SG. In addition, the current study provides high resolution structural information about protein-protein interactions that might be crucial for the SG nucleation (i.e. the structure of the eIF4G1-Pub1 chimera). This chapter highlight the critical role of folded domains in the formation of high order structures and possible in condensates. Both protein-protein recognition between RRMs and eIF4G IDD and self-recognition between RRMs, act together to build up complex tri-dimensional networks that could be the scaffolds of SG cores.

Finally, it is important to highlight that, in our experiments, the formation of condensate involve exclusively protein-protein contacts without needing to add RNA or to apply stressful conditions. Nevertheless, it is clear from previous studies that these two factors play a crucial role on itself. Therefore, we suggest that the protein-protein interactions we described are important for the consolidation of SG cores, in combination with RNA recognition and pH/temperature induced conformational changes.

7.5. MATERIALS AND METHODS

Cloning, protein expression and purification

Plasmids and proteins used in this work are described in the key resources table (**Appendix 3**). DNA fragments corresponding to wild-type constructs of eIF4G1 $_{1-250}$, Pub1 and Pab1 were amplified from *Saccharomyces cerevisiae* genomic DNA using DNA polymerase DNA polymerases KOD or Pfu. These DNA fragments were cloned in a pET28-modified vector that contains: N-terminal thioredoxin A fusion tag, an internal 6xHis tag and a TEV protease site. The mutant in eIF4G1 $_{1-250}$ (eIF4G1 $_{1-250}$ ΔBox1) was obtained with a Quick-change Lightning Kit and specific DNA primers. Plasmids were transformed in *E. coli* BL21 (DE3) competent cells and expressed in kanamycin containing (30 µg/l) LB medium. For isotopic labelling, a K-MOPS derived minimal medium [40] was supplemented with $^{15}NH_4Cl$ (1 g/l) and/or ^{13}C-glucose (4 g/l). In case of amino acid-selective isotope labeling, *Escherichia Coli* autotrophic strains RF6 and RF10 has been used for production of Pro or Lys reverse-labeled eIF4G $_{1-250}$ samples. The K-MOPS minimal medium was supplemented with $^{15}NH_4Cl$ (1 g/l) and/or ^{13}C-glucose (4 g/l) and Proline (0.5 g/l) or Lysine (0.125 g/l), depends of the autotrophic strain used [41]. Cultures of eIF4G1 $_{1-250}$ and eIF4G1 $_{1-250}$ ΔBox1 were grown at 37 ºC until OD_{600} nm = 0.6 – 0.8 when were induced with 0.5 µM of IPTG during 4 hours. Pab1 and Pub1 cultures, after reaching OD_{600} nm = 0.6, were transferred to 25 ºC to induced with IPTG overnight (12 – 16 hours).

For all recombinant proteins, cell pellets were resuspended in lysis buffer (25 mM Potassium phosphate pH 8.0, 300 mM NaCl, 10 mM Imidazole and 1 tablet/50 ml of protease inhibitors cocktail), lysed by sonication and cleared by ultracentrifugation. The supernatant was purified by metal affinity chromatography using HiTrap^{TM} 5ml column and eluted using the buffer 25mM Potassium phosphate pH 8.0, 300 mM NaCl and 300 mM Imidazole. The samples with the fusion protein were exchange to 20 mM Tris pH 8.0 (in case of Pab1 construct this buffer was supplemented with 1 mM DTT), and digested overnight at 4 ºC with homemade TEV protease. In case of Pub1 and Pab1 constructs, the samples were reloaded on the HiTrap nickel column to capture the protease, cleaved fusion protein and uncleaved. The flow through was further purified by ion exchange using an anion exchanger column (Q 5ml) for all the proteins except Pub1 RRM3 that was purified with a cation exchange column (SP 5ml). In either case, proteins were eluted with a linear salt gradient (to 1 M NaCl).

In the case of eIF4G1 $_{1\text{-}250}$ construct we observed that the second nickel column affects negatively to the protein stability and aggregation, therefore we purify the protein away from uncleaved, txA and TEV using a cation exchange column (SP 5ml). Finally proteins were concentrated and buffer exchanged according to their posterior use.

NMR experiments

All samples were prepared in NMR buffer (25 mM Phosphate pH 6.5, 25 mM NaCl, 1mM DTT and 10 % D_2O) and experiments acquired at 25º C on cryoprobe-equipped Bruker AV800 MHz spectrometer. The assignment of the backbone ^1H, ^{15}N and ^{13}C atoms was achieved by following the standard methodology. The 3D experiments 3D HNCO, 3D HNCA, 3D HC(CO)CA [42], 3D HN(CA)CO [43], 3D CBCANH [44], 3D CBCA(CO)NH [45] and 3D (H)CCH-TOCSY which were recorded to assign the side chain resonances [46]. Proteins concentration range between 100 – 200 μM. The chemical shifts of eIF4G1 $_{1\text{-}250}$, Pab1 RRM12 and Pub1 RRM3-eIF4G1 $_{1\text{-}250}$ chimera will be deposited in the Biomagnetic Resonance Database (BMRB). NMR spectra were processed using TOPSPIN v2.1 (Bruker, Inc) and NMRPipe [47] ,and they were analyzed with the program CcpNmr Analysis v2.4.2 [48].

Structure calculations of Pub1-eIF4G $_{1\text{-}250}$ chimera.

Two different Pub1-eIF4G1 chimeras were constructed with the eIF4G$_{37\text{-}51}$ either fused to the N- or C-terminus of Pub1 RRM3. Of these, only the N-terminal fusion proved to have the right topology and its spectra was assigned by triple resonance methods. The structure was determined using a similar protocol to the structure of Pub1 RRM3 [21] using distance restrains from a 2D NOESY (60 ms mixing time) and angular restrains obtained by TALOS+ program [49]. The NMR structure calculation statistics are represented in **Table S1**.

Confocal microscopy and turbidity measurements

Protein labeling. eIF4G1 $_{1\text{-}250}$, eIF4G1 $_{1\text{-}250}$ ΔBox1, Pab1 and Pub1 were purified as described before and stored at 20 ºC until used. The proteins were covalently labeled in the amino groups with Alexa Fluor 488 or Alexa Fluor 647 carboxylic acid succinimidyl ester dyes (Molecular Probes).

For that, the proteins and the probe were incubated 1 h 30 minutes in ice at darkness at [protein : probe] [1 : 3] molar ratio. The protein concentration used to The excess was eliminated by buffer exchange to PBS pH 7.3.

Turbidity measurements. Turbidity of samples containing 1 µM eIF4G1 $_{1-250}$ or eIF4G1 $_{1-250}$ ΔBox1, 5 µM Pub1 and 20 µM Pab1 in presence of 200 g/l Ficoll 70 was determined at room temperature and 350 nm using a Varioskan Flash plate reader (Thermo) The absorbance of these solutions, measured every 10 minutes for 30 minutes, was stable during this time preriod. Reported values, average of 3 – 5 independent measurements ± SD.

Confocal microscopy measurements and data analysis. The pre-condensates generated were visualized in silicone chambers glutted to coverslips. Images were obtained with a Leica TCS SP2 inverted confocal microscope with a HCX PL APO 63x oil immersion objective (N.A. = 1.4-1.6; Leica, Mannheim, Germany). Ar (488 nm) and He-Ne (633 nm) ion lasers were used to excite Alexa 488 and Alexa 467, respectively. Various images were registered depending of different observation fields.

7.6. ACKNOWLEDGMENTS

7.7.   REFERENCES

1.   Aitken CE, Lorsch JR. A mechanistic overview of translation initiation in eukaryotes. Nat Struct Mol Biol. 2012/06/06. 2012;19(6):568–76. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22664984

2.   Preiss T, Hentze MW. From factors to mechanisms: translation and translational control in eukaryotes. Curr Opin Genet Dev. 1999/10/06. 1999;9(5):515–21. Available from: https://www.ncbi.nlm.nih.gov/ pubmed/ 10508691

3.   Sonenberg N, Hinnebusch AG. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell. 2009/02/26. 2009;136(4):731–45. Available from: https://www.ncbi.nlm.nih. gov/pubmed /19239892

4.   Dever TE, Kinzy TG, Pavitt GD. Mechanism and Regulation of Protein Synthesis in. 2016;203(May):65–107.

5.   Tarun Jr. SZ. Translation initiation factor eIF4G mediates in vitro poly (A) tail-dependent translation. 1997;94(August):9046–51.

6.   Wells SE, Hillner PE, Vale RD, Sachs AB. Circularization of mRNA by eukaryotic translation initiation factors. Mol Cell. 1998/08/14. 1998;2(1):135–40. Available from: https://www.ncbi.nlm.nih.gov /pubmed/9702200

7.   Buchan JR, Muhlrad D, Parker R. P bodies promote stress granule assembly in Saccharomyces cerevisiae. J Cell Biol. 2008/11/05. 2008;183(3):441–55. Available from: https://www.ncbi.nlm.nih.gov/pubmed /18981231

8.   Buchan JR, Yoon JH, Parker R. Stress-specific composition, assembly and kinetics of stress granules in Saccharomyces cerevisiae. J Cell Sci. 2010/12/22. 2011;124(Pt 2):228–39. Available from: https:// www.ncbi. nlm.nih.gov/pubmed/21172806

9.   Hoyle NP, Castelli LM, Campbell SG, Holmes LE, Ashe MP. Stress-dependent relocalization of translationally primed mRNPs to cytoplasmic granules that are kinetically and spatially distinct from P-bodies. J Cell Biol. 2007/10/03. 2007;179(1):65–74. Available from: https://www.ncbi.nlm.nih.gov/ pubmed/17908917

10.   Hofweber M, Dormann D. Friend or foe-Post-translational modifications as regulators of phase separation and RNP granule dynamics. J Biol Chem. 2018/12/28. 2019;294(18):7137–50. Available from: https:// www.ncbi.nlm.nih.gov/pubmed/30587571

11.   Kroschwald S, Munder MC, Maharana S, Franzmann TM, Richter D, Ruer M, et al. Different Material States of Pub1 Condensates Define Distinct Modes of Stress Adaptation and Recovery. Cell Rep. 2018/06/14. 2018;23(11):3327–39. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29898402

12.   Das S, Das B. eIF4G—an integrator of mRNA metabolism? FEMS Yeast Res . 2016 Oct 14;16(7). Available from: https://doi.org/10.1093/femsyr/fow087

13.   Costello J, Castelli LM, Rowe W, Kershaw CJ, Talavera D, Mohammad-Qureshi SS, et al. Global mRNA selection mechanisms for translation initiation. Genome Biol. 2015 Jan 5;16(1):10. Available from: https://www. ncbi.nlm.nih.gov/pubmed/25650959

14. Safaee N, Kozlov G, Noronha AM, Xie J, Wilds CJ, Gehring K. Interdomain allostery promotes assembly of the poly(A) mRNA complex with PABP and eIF4G. Mol Cell. 2012/10/09. 2012;48(3):375–86. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23041282

15. Plank T-DM, Kieft JS. The structures of nonprotein-coding RNAs that drive internal ribosome entry site function. Wiley Interdiscip Rev RNA. 2012/01/03. 2012;3(2):195–212. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22215521

16. Anderson JT, Paddy MR, Swanson MS. PUB1 Is a Major Nuclear and Cytoplasmic Polyadenylated RNA-Binding Protein in Saccharomyces cerevisiae. 1993;13(10):6102–13.

17. Duttagupta R, Tian B, Wilusz CJ, Khounh DT, Soteropoulos P, Ouyang M, et al. Global analysis of Pub1p targets reveals a coordinate control of gene expression through modulation of binding and stability. Mol Cell Biol. 2005 Jul;25(13):5499–513. Available from: https://www.ncbi.nlm.nih.gov/pubmed/15964806

18. Costello JL, Kershaw CJ, Castelli LM, Talavera D, Rowe W, Sims PFG, et al. Dynamic changes in eIF4F-mRNA interactions revealed by global analyses of environmental stress responses. Genome Biol . 2017 Oct 27;18(1):201. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29078784

19. Riback JA, Katanski CD, Kear-Scott JL, Pilipenko E V, Rojek AE, Sosnick TR, et al. Stress-Triggered Phase Separation Is an Adaptive, Evolutionarily Tuned Response. Cell. 2017/03/12. 2017;168(6):1028-1040 e19. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28283059

20. Eleanor J. Taylor †, Susan G. Campbell ‡, Christian D. Griffiths PJR, John W. Slaven,* Richard J. Harrison,§ Paul F.G. Sims 储, Pavitt GD, Daniela Delneri and MPA. Fusel Alcohols Regulate Translation Initiation by Inhibiting eIF2B to Reduce Ternary Complex in a Mechanism That May Involve Altering the Integrity and Dynamics of the eIF2B Body Eleanor. Mol Biol Cell. 2010;21(22):4042–56.

21. Santiveri CM, Mirassou Y, Rico-Lastres P, Martinez-Lumbreras S, Perez-Canadillas JM. Pub1p C-terminal RRM domain interacts with Tif4631p through a conserved region neighbouring the Pab1p binding site. PLoS One. 2011/09/21. 2011;6(9):e24481. Available from: https://www.ncbi.nlm.nih.gov/pubmed /21931728

22. Martinez-Lumbreras S. The role of Gbp2p , Nab2p and Pub1p along the mRNA cycle : structural and molecular recognition studies by NMR and other biophysical techniques (Doctoral thesis). Univ Autónoma Madrid. 2013;

23. Rasia RM, Brutscher B, Plevin MJ. Selective isotopic unlabeling of proteins using metabolic precursors: application to NMR assignment of intrinsically disordered proteins. Chembiochem. 2012;13(5):732–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22408059

24. Golovanov AP, Blankley RT, Avis JM, Bermel W. Isotopically discriminated NMR spectroscopy: A tool for investigating complex protein interactions in vitro. J Am Chem Soc. 2007;129(20):6528–35.

25. Kessler SH, Sachs AB. RNA recognition motif 2 of yeast Pab1p is required for its functional interaction with eukaryotic translation initiation factor 4G. Mol Cell Biol. 1998;18(1):51–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9418852%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC121449

26.   Tarun Jr. SZ, Wells SE, Deardorff JA, Sachs AB. Translation initiation factor eIF4G mediates in vitro poly(A) tail-dependent translation. Proc Natl Acad Sci U S A. 1997/08/19. 1997;94(17):9046–51. Available from: https://www.ncbi.nlm.nih.gov/pubmed/9256432

27.   Melamed D, Young DL, Miller CR, Fields S. Combining Natural Sequence Variation with High Throughput Mutational Data to Reveal Protein Interaction Sites. 2015;1–21.

28.   Imataka H, Gradi A, Sonenberg N. A newly identified N-terminal amino acid sequence of human eIF4G binds poly(A)-binding protein and functions in poly(A)-dependent translation. EMBO J. 1998; 17(24):7480–9.

29.   Park EH, Walker SE, Lee JM, Rothenburg S, Lorsch JR, Hinnebusch AG. Multiple elements in the eIF4G1 N-terminus promote assembly of eIF4G1*PABP mRNPs in vivo. EMBO J . 2010/12/09. 2011;30(2):302–16. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21139564

30.   Gilbert W V, Zhou K, Butler TK, Doudna JA. Cap-Independent Translation Is Required for Starvation-Induced Differentiation in Yeast. Science (80). 2007 Aug 31;317(5842):1224 LP – 1227. Available from: http://science.sciencemag.org/content/317/5842/1224.abstract

31.   Rachfall N, Heinemeyer I, Morgenstern B, Valerius O, Braus GH. 5'TRU: identification and analysis of translationally regulative 5'untranslated regions in amino acid starved yeast cells. Mol Cell Proteomics. 2011/03/28. 2011 Jun ;10(6) :M110.003350-M110.003350. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21444828

32.   Cléry A, Sinha R, Anczuków O, Corrionero A, Moursy A, Daubner GM, et al. Isolated pseudo-RNA-recognition motifs of SR proteins can regulate splicing using a noncanonical mode of RNA recognition. Proc Natl Acad Sci U S A. 2013/07/08. 2013 Jul 23;110(30):E2802–11. Available from: https://www.ncbi.nlm.nih.gov /pubmed/23836656

33.   Martínez-Lumbreras S, Taverniti V, Zorrilla S, Séraphin B, Pérez-Cañadillas JM. Gbp2 interacts with THO/TREX through a novel type of RRM domain. Nucleic Acids Res. 2015/11/23. 2016 Jan 8;44(1):437–48. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26602689

34.   Ngo JCK, Giang K, Chakrabarti S, Ma C-T, Huynh N, Hagopian JC, et al. A sliding docking interaction is essential for sequential and processive phosphorylation of an SR protein by SRPK1. Mol Cell . 2008 Mar 14;29(5):563–76. Available from: https://www.ncbi.nlm.nih.gov/pubmed/18342604

35.   Otero LJ, Ashe MP, Sachs AB. The yeast poly(A)-binding protein Pab1p stimulates in vitro poly(A)-dependent and cap-dependent translation by distinct mechanisms. EMBO J. 1999 Jun 1;18(11):3153–63. Available from: https://www.ncbi.nlm.nih.gov/pubmed/10357826

36.   Berset C, Zurbriggen A, Djafarzadeh S, Altmann M, Trachsel H. RNA-binding activity of translation initiation factor eIF4G1 from Saccharomyces cerevisiae. RNA. 2003 Jul;9(7):871–80. Available from: https://www.ncbi.nlm.nih.gov/pubmed/12810920

37.   Lin Y, Protter DS, Rosen MK, Parker R. Formation and Maturation of Phase-Separated Liquid Droplets by RNA-Binding Proteins. Mol Cell. 2015/09/29. 2015;60(2):208–19. Available from: https:// www.ncbi.nlm.nih.gov /pubmed/26412307

38. Jain S, Wheeler JR, Walters RW, Agrawal A, Barsic A, Parker R. ATPase-Modulated Stress Granules Contain a Diverse Proteome and Substructure. Cell. 2016/01/19. 2016;164(3):487–98. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26777405

39. Preiss T, Hentze MW. Dual function of the messenger RNA cap structure in poly(A)-tail-promoted translation in yeast. Nature. 1998;392(6675):516–20. Available from: https://doi.org/10.1038/33192

40. Neidhardt FC, Bloch PL, Smith DF. Culture medium for enterobacteria. J Bacteriol. 1974;119(3):736–47.

41. Lin MT, Fukazawa R, Miyajima-nakano Y, Matsushita S, Choi SK, Iwasaki T, et al. Escherichia coli Auxotroph Host Strains for Amino Acid-Selective Isotope Labeling of Recombinant Proteins. 1st ed. Vol. 565, Isotope labeling of biomolecules labeling Methods. Elsevier Inc.; 2015. 45–66 p. Available from: http://dx.doi.org/10.1016/bs.mie.2015.05.012

42. Grzesiek S, Bax A. Improved 3D triple-resonance NMR techniques applied to a 31 kDa protein. J Magn Reson. 1992;96(2):432–40.

43. Clubb RT, Thanabal V, Wagner G. A constant-time three-dimensional triple-resonance pulse scheme to correlate intraresidue 1HN, 15N, and 13C' chemical shifts in 15N13C-labelled proteins. J Magn Reson. 1992;97(1):213–7.

44. Grzesiek S, Bax A. An efficient experiment for sequential backbone assignment of medium-sized isotopically enriched proteins. J Magn Reson. 1992;99(1):201–7.

45. Grzesiekt S, Bax A. Correlating Backbone Amide and Side Chain Resonances in Larger Proteins by Multiple Relayed Triple Resonance NMR. J Am Chem Soc. 1992;114(16):6291–3.

46. Sattler M, Schleucher J, Griesinger C. Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. Prog Nucl Magn Reson Spectrosc. 1999;34(2):93–158.

47. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. NMRPipe: A multidimensional spectral processing system based on UNIX pipes. J Biomol NMR. 1995;6(3):277–93. Available from: http://www.scopus.com/inward/record.url?eid=2

48. Skinner SP, Fogh RH, Boucher W, Ragan TJ, Mureddu LG, Vuister GW. CcpNmr AnalysisAssign : a flexible platform for integrated NMR analysis. J Biomol NMR. 2016;66(2):111–24.

49. Shen Y, Delaglio F, Cornilescu G, Bax A. TALOS+: A hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR. 2009;44(4):213–23.

7.8. SUPPORTING INFORMATION
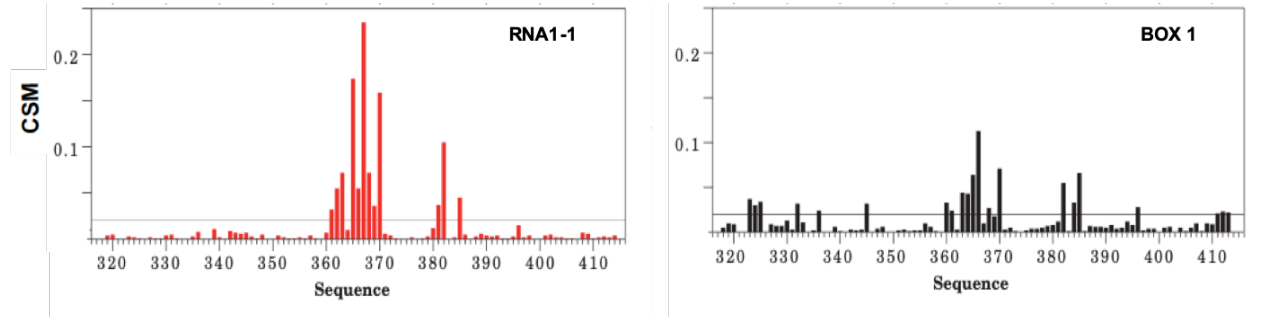
Figure S1. Chemical shift mapping (CSM) of Pub1 RRM3 in its titration with eIF4G1 $_{1-250}$ peptides (RNA1_1 in red and BOX1 in black).
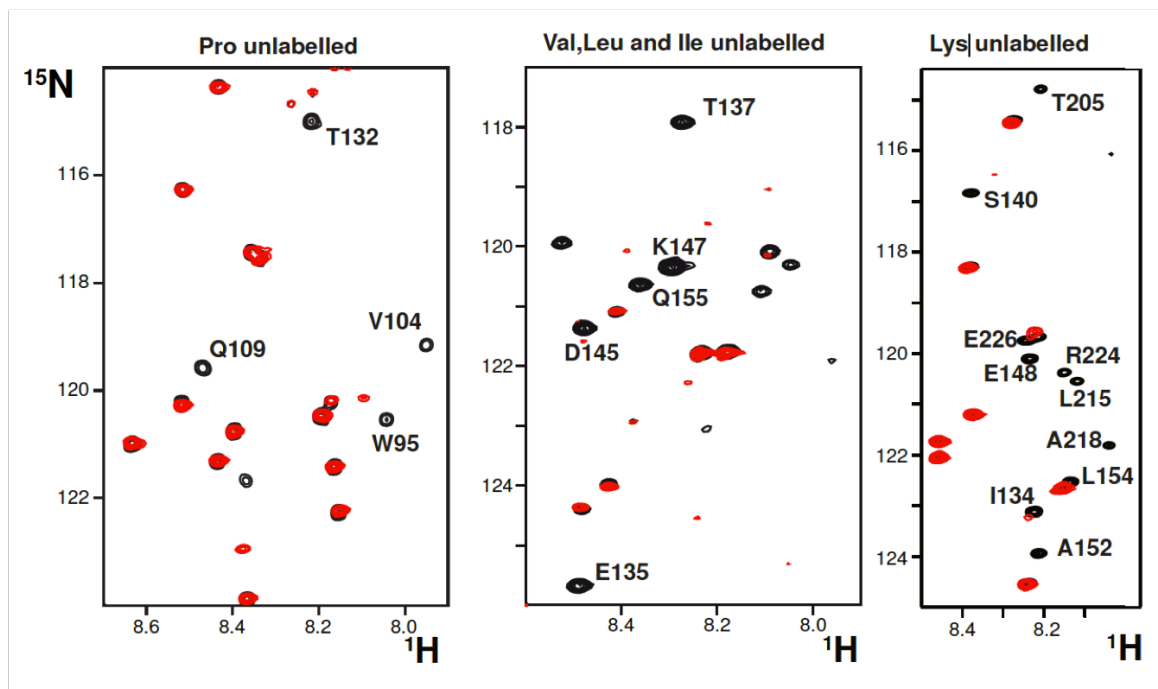
Figure S2. Isotopically discriminated ^{1}H-^{15}N correlation spectra (IDIS). eIF4G1 $_{1-250}$ reverse Pro and Lys samples were expressed using *Escherichia Coli* auxotrophic strains RF6 and RF10 in ^{13}C/^{15}N minimal media supplemented with unlabelled amino acids. The standard *Escherichia Coli* BL21 (i.e. ^{12}CO) and therefore map residues following Pro, Lys, Val, Ile and Leu. IDIS spectra of the complex with Pab1 RRM12 (unlabelled).
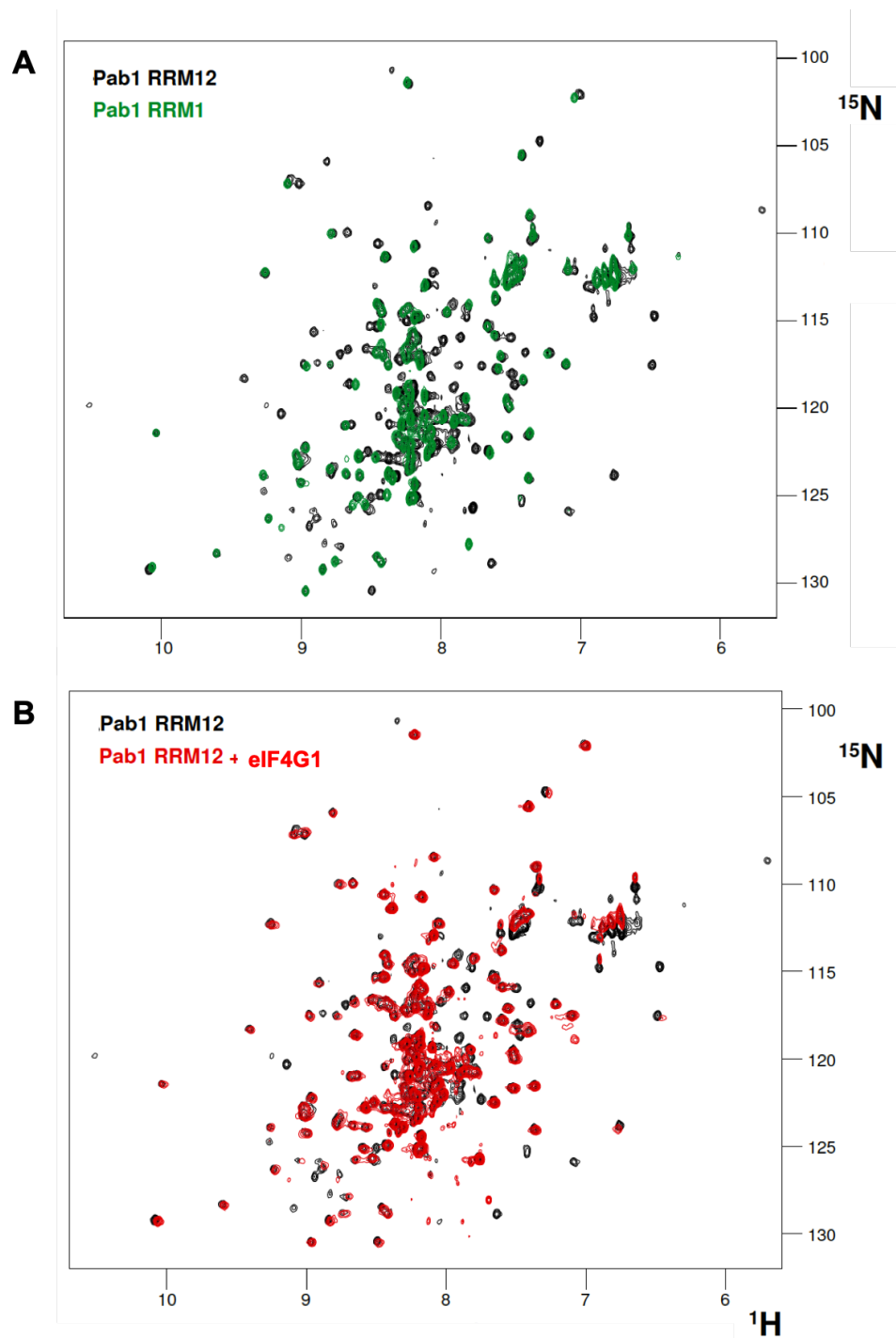
Figure S3. (A) Superposition of the NMR spectra of Pab1 RRM1 (green) and Pab1 RRM12 (black). Disordered region preceding the RRM1 domains is included in both construct. **(B)** 2D TROSY spectra of the complex. Both spectra were acquired with 15N-labelled Pab1 RRM12 at concentration of 100 µM but the 2D TROSY spectra was recoded with 10-times more scans.
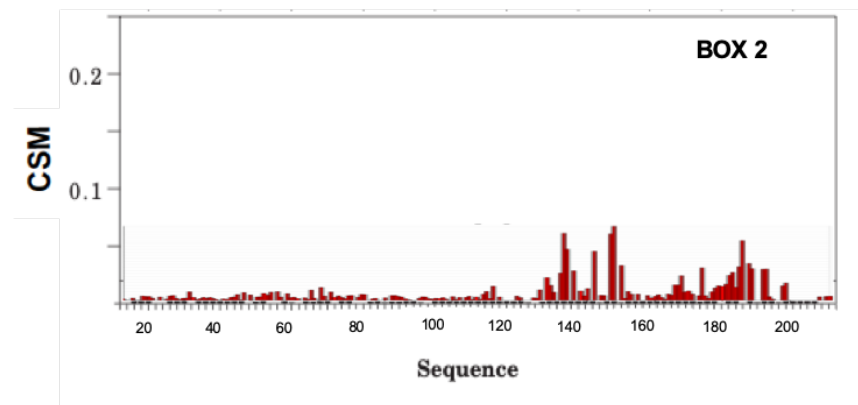
Figure S4. Chemical shift mapping (CSM) of Pab1 RRM12 in its titration with eIF4G1 $_{1-250}$ peptide BOX2.
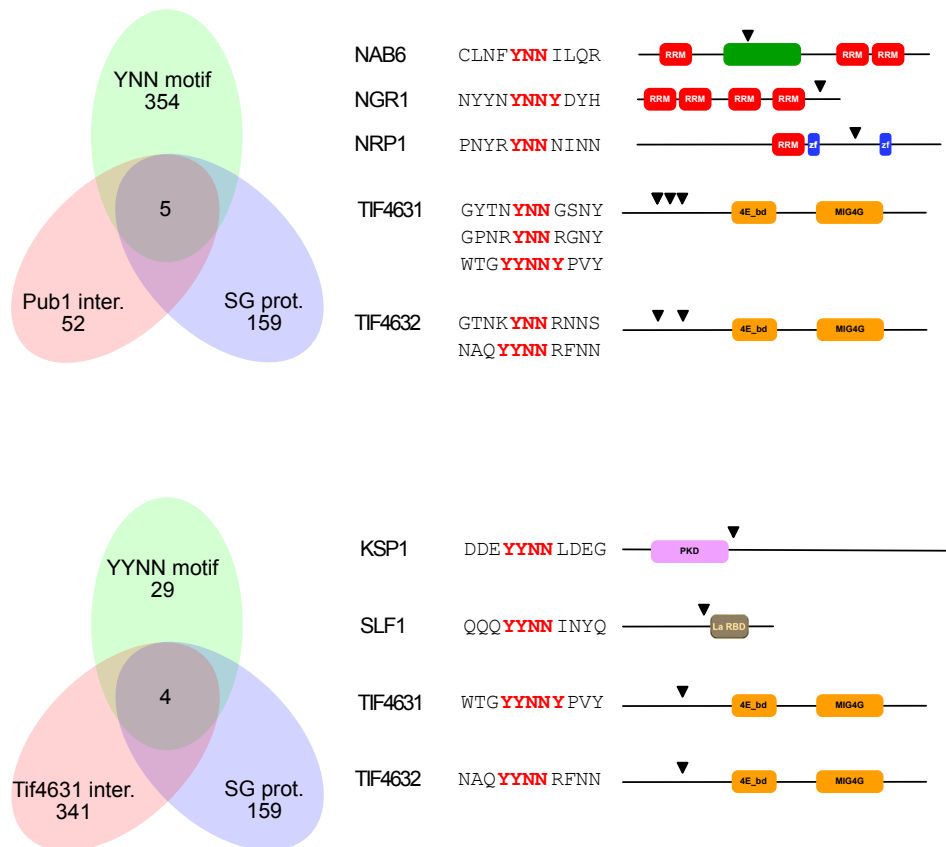
Figure S5. Analysis of the YNN and YYNN motifs on yeast proteome. The Venn diagrams shows the intersection of total occurrences (green), with SG components (blue) [38] and Pub1/eIF4G interacting proteins (red) (SGD: https://www.yeastgenome.org/). The proteins meeting the three criteria are represented on the right with the sequences in which the motifs are embedded and their location in the sequence (down black triangles).

eIF4G1-Alexa 488

eIF4G1-Alexa 488
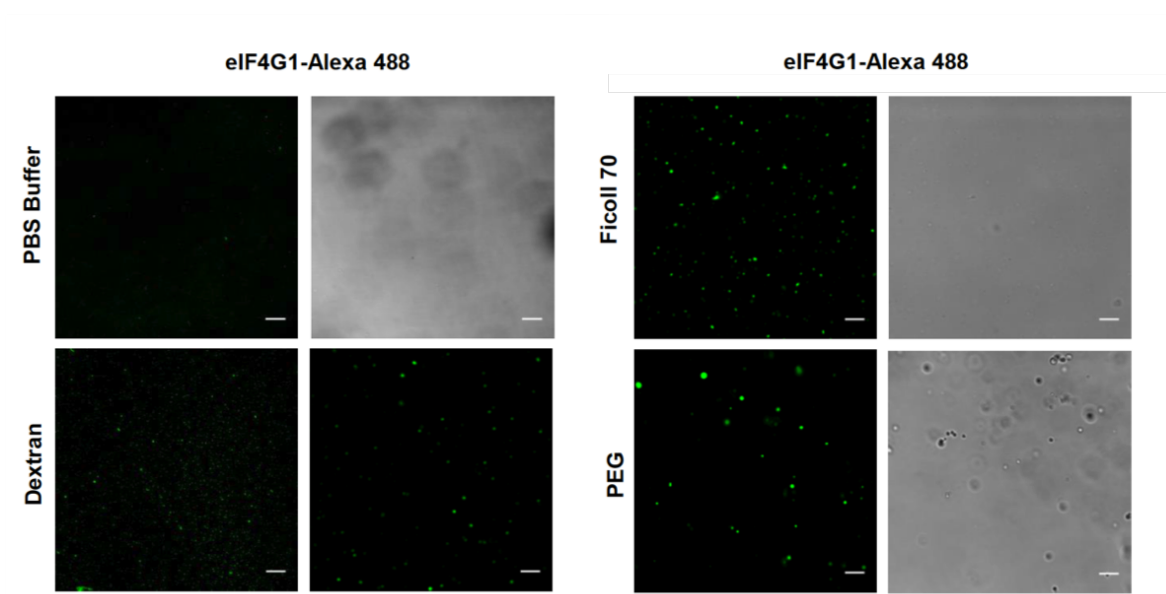
Figure S6. Representative confocal images of condensates of the eIF4G1 $_{1-250}$, Pub1 and Pab1 complex condensates formed in 200 g/l dextran 500, Ficoll 70 or PEG, and absence of condensates in dilute solution (PBS). Scale bars: 20 μm.

eIF4G1-Alexa 488   **Pab1-Alexa 647**
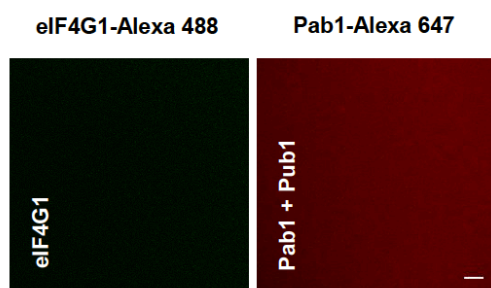
Figure S7. Representative confocal images of no condensates of the eIF4G1 $_{1-250}$, Pub1 and Pab1 proteins in 200 g/l Ficoll 70. Scale bars: 20 μm.

Table S1: NMR Structure calculation statistics

NMR experimental restraints	
NOE-derived	1298
Intraresidue	347
Sequential	284
Medium range (1< i-j < 4)	191
Long-range (1< i-j < 4)	476
eIF4G-Pub1	26
Structure statistics (20 structures)	
Cyana target function	4.77
Distance violations	11 ± 0.3
Maximun distance violations (Å)	0.53
Average RMSD	
Backbone	0.61 ± 0.8
Heavy	1.03 ± 0.8

CONCLUSIONS

The last decade has witnessed a sharp surge in the number of experimentally identified intrinsically disordered domains (IDDs) and proteins (IDPs), along with a steady advance in the accuracy of tools that predict protein disorder. It has been convincingly demonstrated in a large number of examples, such as C-H1.0 and eIF4G1 studied in this thesis, that the IDPs are very important to protein function, such as gene expression regulation. Novel experimental methods and strategies, especially in NMR spectroscopy, have been devised and successfully employed to characterize these proteins. With a better understanding of the conformational properties and resultant functional insights of the IDPs studied in the previous chapters, the following can be concluded.

o   A novel ^{13}C-detected assignment strategy based on a set of only four NMR experiments was validated by its application to the C-terminal intrinsically domain of Histone H1.0 (C-H1.0). The strategy is useful for assigning and analyzing Pro-rich IDPs with repetitive sequence.

o   Based on chemical shift deviations and heteronuclear ^{1}H,^{15}N NOEs, C-H1.0 was shown to be highly disordered both non-phosphorylated and triphosphorylated (pT-C-H1.0) forms. Consequently, the regulatory effect of phosphorylation on the C-H1.0 function does not appear to depend on the acquisition of secondary structure.

o   Upon DNA binding C-H1.0 and p-T-C-H1.0 remain as highly disordered as in their free states, which confirms that IDPs and IDDs can remain disordered when bound to their target partners.

o   DNA interaction occurs mainly by the C-terminal region of C-H1.0, being unrelated to the distribution of Lys residues along the C-H1.0 sequence. Phosphorylation promotes a decrease in DNA interaction, maybe by altering the charge balance. In the phosphorylated DNA-bound p-T-CH1.0, the conformational ensemble around the middle phosphorylation motif seems to suffer a change relative to the free domain. This local change together with the weaker DNA-affinity might be important for chromatin de-condensation triggered by phosphorylation.

o   The structural study of two model C-H1.0-derived peptides containing a single phosphorylation motif, both non-phosphorylated and phosphorylated, has shown that they are mainly random coil in aqueous solution, behaving as the full-length C-H1.0, and formed α-helical structures in the presence of 90 % trifluoroethanol (TFE). These results indicate that minimalist approaches based on model peptides are useful to obtain structural data of IDPs difficult to study.

o The application of selective labelling methods greatly facilitates the assignment of large IDPs as eIF4G1 $_{1-250}$ and the study of their interactions. A novel strategy was used combining these experimental data with knowledge-based π-cation and π-π to calculate a eIF4G1 $_{1-250}$ structural ensemble (158 structures) that faithfully reproduce the PRE data. The ensemble is structurally diverse apart from a highly populated α-helix in the BOX3 element that interacts transiently with RNA1 and BOX1 elements, and these two among themselves. These transient contacts are highly variable and explain the unusual stability of the protein. The eIF4G1 $_{1-250}$ structure represents one of the first examples of an IDP structure.

o The structural basis of Pab1 RRM12, Pub1 RRM3 and eIF4G1 $_{1-250}$ self-recognition enables them to be multivalent and can give rise to extensive oligomeric networks. eIF4G1 $_{1-250}$ can dimerize or oligomerize through contacts involving the BOX1 and RNA1 domains. Pab1 RRM12 and Pub1 RRM3 can form dimers and circular species through interactions involving their RRMs and short sequences in their IDD. Pab1 and Pub1 do not interact with each other.

o The NMR study of the interaction between eIF4G1 $_{1-250}$ and Pab1 shows a new binding site, which could play a new role in the cap-independent translation initiation. Simultaneous binding of Pab1 RRM12, Pub1 RRM3 and eIF4G1 $_{1-250}$ induce aggregation, probably by uncovering the amyloid-like properties of the prion-like sequence in BOX1. Mixtures of eIF4G/Pab1 and eIF4G/Pab1/Pub1 form micrometer size protein condensates that could be detected by fluorescence confocal microscopy.

o A model involving different types of multivalent interactions has been proposed. This model could resemble some of the structural characteristic of stress granule cores.

CONCLUSIONES

Durante la última década el número de dominios y proteínas intrínsecamente desordenadas (IDDs o IDPs) identificadas experimentalmente ha aumentado considerablemente. A ello ha contribuido los avances constantes en las metodologías computacionales que las predicen. Numerosos ejemplos, entre los que se encuentran los dos estudiados en esta tesis (C-H1.0 y eIF4G), reflejan la gran importancia de tener dominios desordenados para cumplir funciones proteicas como la regulación de la expresión génica. Se han desarrollado nuevas metodologías y estrategias experimentales, especialmente en el campo de la espectroscopia de RMN, que permiten la caracterización de estas IDPs de forma satisfactoria. Gracias al conocimiento adquirido en esta tesis acerca de las propiedades conformacionales y las funciones que llevan a cabo las proteínas C-H1.0 y eIF4G, se puede concluir lo siguiente:

o Se ha validado una nueva estrategia de asignación de RMN basada en la detección del ^{13}C a través de sólo cuatro experimentos mediante al dominio C-terminal de la Histona H1.0 (C-H1.0). Esta estrategia es útil para asignar y analizar IDPs de secuencias repetitivas y ricas en prolinas.

o Basándose en las desviaciones de desplazamiento químico y los NOEs heteronucleares ^{1}H,^{15}N, se ha demostrado que C-H1.0 es una proteína altamente desordenada tanto en su forma no fosforilada como en la trifosforilada (pT-C-H1.0). Como consecuencia, la regulación de la función de C-H1.0 por la fosforilación no se debe a la formación de estructura secundaria.

o Al unirse al ADN, las proteínas C-H1.0 y pT-C-H1.0 permanecen altamente desordenadas, lo que confirma que algunas IDPs o IDDs pueden unirse a su diana manteniéndose desordenadas.

o La interacción con el ADN ocurre principalmente por la región C-terminal de C-H1.0 y no está relacionada con la distribución de Lys a lo largo de la secuencia de C-H1.0. La fosforilación disminuye esta interacción con el ADN, lo cual podría deberse a la alteración en el equilibrio de cargas. Cuando pT-C-H1.0 se une al ADN, el motivo de fosforilación central de la secuencia parece sufrir más cambios con respecto al dominio libre. Este cambio local junto a una menor afinidad por el ADN podría ser importante para la de-condensación de la cromatina desencadenada por la fosforilación.

o El estudio estructural de péptidos de la proteína C-H1.0 con un único motivo de fosforilación (fosforilado y no fosforilado) ha demostrado que carecen de estructura en solución acuosa,

comportándose igual que en el dominio completo C-H1.0. Sin embargo, en presencia de trifluoroetanol (TFE) al 90 % forman hélices α. Estos resultados indican que estrategias minimalistas usando péptidos modelo son útiles para obtener datos estructurales de IDPs difíciles de estudiar.

o El uso de métodos de etiquetado selectivo de amino ácidos facilita la asignación de IDPs como eIF4G1 1-250 y el estudio de sus interacciones. Se ha utilizado una nueva estrategia que combina los datos experimentales estructurales con el conocimiento acerca de interacciones π-cation y π-π para calcular un conjunto representativo de estructuras para eIF4G1 1-250 (158 estructuras), que además reproducen los datos de PRE. Las estructuras de este conjunto son muy diversas, pero la mayoría contienen una hélice en la región BOX3 que interactúa de forma transitoria con los dominios RNA1 y BOX1, que a su vez interaccionan entre sí. Los contactos transitorios son muy variables y explican la baja estabilidad de la proteína. La estructura de eIF4G1 1-250 es uno de los primeros ejemplos de estructura de IDP.

o La capacidad de auto-reconocimiento les permite a Pab1 RRM12, Pub1 RRM3 y eIF4G1 1-250 ser multivalentes y dar lugar a extensas redes de oligómeros. eIF4G1 1-250 se puede dimerizar u oligomerizar a través de contactos entre los dominios BOX1 y RNA1. Además, Pab1 RRM12 y Pub1 RRM3 pueden formar dímeros y especies circulares mediante interacciones entre sus dominios RRM y secuencias de su parte desordenada. Por otro lado, Pab1 y Pub1 no interaccionan entre sí.

o El estudio por RMN de la interacción entre eIF4G1 1-250 y Pab1 permite identificar un nuevo sitio de unión que podría tener un papel novedoso en el inicio de la traducción. La unión simultánea de Pab1 RRM12, Pub1 RRM3 y eIF4G1 1-250 induce agregación probablemente debido a las propiedades amiloides de la secuencia del dominio BOX1 de eIF4G1 1-250. Las mezclas eIF4G/Pab1 y eIF4G/Pab1/Pub1 dan lugar a condensados proteicos de tamaño micrométrico que se detectan por microscopía confocal de fluorescencia.

o Se ha propuesto un modelo que involucra distintos tipos de interacciones multivalentes. Este modelo podría representar las características estructurales de los núcleos de los gránulos de estrés.

APPENDIXES

I. APPENDIX

Assignment of non-phosphorylated C-terminal domain of histone H1.0

	AA	$^{13}C_\alpha$	$^{13}C_\beta$	$^{13}C'$	^{15}N	$^1H^N$		AA	$^{13}C_\alpha$	$^{13}C_\beta$	$^{13}C'$	^{15}N	$^1H^N$
1	M	55.26	33.31	173.27	-	-	30	K	56.40	33.25	176.73	122.44	-
2	D	54.67	41.17	175.52	123.93	-	31	K	56.38	33.18	176.22	123.32	-
3	E	54.77	29.78	174.61	122.94	-	32	A	52.43	19.23	177.43	126.44	8.43
4	P	63.37	32.05	177.03	137.54	-	33	A	52.62	19.19	177.89	124.11	8.39
5	K	56.56	32.20	176.87	121.50	-	34	S	58.34	64.10	174.46	115.76	-
6	R	56.33	30.97	176.42	122.46	-	35	K	56.23	33.10	175.98	123.60	8.40
7	S	58.44	63.98	174.73	117.40	-	36	A	50.59	18.00	175.54	127.23	8.39
8	V	62.42	32.91	175.84	121.62	8.19	37	P	63.07	32.09	177.06	135.82	-
9	A	52.64	19.10	177.41	126.94	8.28	38	S	58.51	64.09	174.67	116.69	8.50
10	F	57.90	39.62	175.59	119.82	8.13	39	K	56.36	33.08	176.32	123.70	8.44
11	K	56.23	33.21	176.04	123.55	8.15	40	K	54.30	32.38	174.54	124.43	8.40
12	K	56.53	33.10	176.76	123.32	8.35	41	P	63.14	32.15	176.85	137.41	-
13	T	62.27	70.45	174.40	116.45	8.26	42	K	56.36	33.22	176.37	122.22	-
14	K	56.39	33.07	176.33	124.46	8.41	43	A	52.20	19.23	177.55	126.02	8.41
15	K	56.60	33.04	176.35	123.64	8.50	44	T	60.07	69.81	172.93	117.32	8.31
16	E	56.41	30.49	176.26	123.48	-	45	P	63.23	32.12	176.83	139.28	-
17	V	62.33	32.80	176.02	123.00	8.34	46	V	62.65	32.91	176.23	121.40	-
18	K	56.36	33.05	176.28	126.11	8.44	47	K	56.40	33.03	176.34	126.18	8.39
19	K	56.41	33.07	176.39	124.15	8.44	48	K	56.33	33.27	176.17	124.01	8.42
20	V	62.14	33.08	175.68	122.56	8.25	49	A	52.30	19.28	177.62	126.29	8.40
21	A	52.34	19.27	177.49	128.79	8.49	50	K	56.41	33.04	176.58	121.85	8.40
22	T	59.92	69.76	172.94	117.10	8.30	51	K	56.34	33.15	176.34	123.66	8.41
23	P	63.37	32.20	176.89	139.27	-	52	K	-	-	174.54	125.09	8.50
24	K	56.56	33.08	176.73	122.41	-	53	P	63.08	32.16	176.58	137.42	-
25	K	56.36	33.09	176.14	123.21	-	54	A	52.32	19.25	177.47	124.81	-
26	A	52.33	19.26	177.27	126.24	-	55	A	52.19	19.19	177.66	123.92	8.36
27	A	52.25	19.26	177.54	124.35	8.34	56	T	60.01	69.86	172.95	116.81	8.28
28	K	54.31	32.43	174.62	122.48	8.37	57	P	63.20	32.18	176.87	139.34	-
29	P	63.26	32.37	176.89	137.35	-	58	K	56.46	33.10	176.70	122.41	-

59	K	56.34	33.14	176.16	123.33	-	80	K	56.54	33.02	176.79	121.61	8.43
60	A	52.33	19.28	177.59	126.29	-	81	T	62.00	70.02	174.31	117.21	8.31
61	K	56.41	33.10	176.47	121.82	-	82	V	62.22	32.84	175.83	124.06	8.33
62	K	54.30	32.54	174.50	124.66	-	83	K	54.19	32.54	174.40	127.50	8.50
63	P	63.17	32.14	176.67	137.25	-	84	P	63.17	32.20	176.86	137.17	-
64	K	56.53	32.89	176.56	122.46	-	85	K	56.51	33.00	176.53	122.24	-
65	V	62.43	32.87	175.89	123.20	8.28	86	A	52.71	19.23	177.81	126.03	8.40
66	V	62.22	32.79	175.83	126.09	8.38	87	K	56.65	32.87	176.91	121.46	8.45
67	K	56.36	33.03	176.21	126.68	8.49	88	S	58.48	64.10	174.84	117.54	-
68	V	62.32	32.88	175.76	122.77	8.25	89	S	58.43	63.95	174.48	118.43	-
69	K	54.29	32.46	174.43	127.50	8.47	90	A	52.76	19.03	177.88	126.23	8.35
70	P	62.99	32.03	176.84	137.08	-	91	K	56.52	32.68	176.83	120.85	-
71	V	62.26	32.89	176.28	121.18	-	92	R	56.16	30.81	176.19	122.78	8.38
72	K	56.40	33.07	176.05	125.92	8.43	93	A	52.53	19.22	177.87	125.79	8.43
73	A	52.45	19.22	177.65	126.56	8.44	94	S	58.75	64.15	174.79	115.80	8.37
74	S	58.22	64.15	174.20	116.35	8.39	95	K	56.56	32.66	176.65	123.81	-
75	K	54.44	32.51	174.50	124.54	8.41	96	K	-	-	-	122.80	8.34
76	P	63.26	32.18	176.87	137.42	-	97	K	-	-	-	-	-
77	K	56.45	33.09	176.72	122.49	-	98	R	-	-	-	-	-
78	K	56.45	33.28	176.19	123.42	-	99	S	-	-	-	-	-
79	A	52.44	19.34	177.64	126.46	8.44	100	H	-	-	-	-	-

Appendix 1. The sample was prepared at 1 mM protein concentration in H_2O/D_2O 9:1 v/v at 302.1K and pH 5.5 conditions. (-) Not observed due to fast HN <-> H_2O fast exchange or signal overlap.

II.   APPENDIX

Assignment of phosphorylated C-terminal domain of histone H1.0

	AA	$^{13}C_\alpha$	$^{13}C_\beta$	$^{13}C'$	^{15}N	$^1H^N$		AA	$^{13}C_\alpha$	$^{13}C_\beta$	$^{13}C'$	^{15}N	$^1H^N$
1	M	55.24	32.76	172.06	-	-	30	K	56.45	33.02	176.65	122.32	8.38
2	D	54.45	41.27	175.33	123.98	8.81	31	K	56.35	33.17	176.18	123.04	8.29
3	E	54.38	29.75	174.41	122.71	8.48	32	A	52.64	19.08	177.40	126.26	8.36
4	P	63.52	32.17	176.94	137.32	-	33	A	52.80	19.06	177.83	123.97	8.32
5	K	56.36	32.61	-	121.56	8.28	34	S	58.50	64.16	174.39	115.51	8.26
6	R	56.26	31.06	176.33	-	-	35	K	56.19	33.19	175.92	123.45	8.31
7	S	58.35	63.61	174.67	117.32	8.35	36	A	50.49	17.79	175.51	127.03	8.32
8	V	62.37	32.78	175.80	121.47	-	37	P	63.36	32.00	177.01	135.76	-
9	A	52.59	19.19	177.38	126.79	-	38	S	58.59	64.15	174.56	116.41	8.39
10	F	57.92	39.76	175.58	119.69	-	39	K	56.32	32.95	176.22	123.52	8.34
11	K	55.98	33.42	176.05	123.33	-	40	K	54.37	32.41	174.51	124.33	8.31
12	K	56.53	33.10	176.70	123.11	8.33	41	P	63.23	32.36	176.71	137.33	-
13	T	61.56	70.06	174.35	116.05	-	42	K	56.24	33.04	176.17	121.86	8.35
14	K	56.08	33.21	176.28	124.22	8.36	43	A	52.50	19.36	177.32	126.14	8.33
15	K	56.07	32.92	176.29	123.48	8.35	44	T	59.77	73.41	172.26	117.96	8.55
16	E	56.04	30.70	176.16	123.29	8.44	45	P	63.53	32.10	176.91	139.67	-
17	V	62.22	30.85	175.93	122.75	8.26	46	V	63.02	31.99	176.34	122.01	8.31
18	K	56.27	33.16	176.14	126.02	8.38	47	K	56.36	33.08	176.34	126.20	8.43
19	K	56.42	33.27	176.28	124.18	8.37	48	K	56.43	33.15	176.13	123.44	8.30
20	V	62.08	32.70	175.59	122.60	8.19	49	A	52.58	19.25	177.57	125.90	8.31
21	A	52.37	19.27	177.28	128.70	8.40	50	K	56.33	33.19	176.51	121.60	8.31
22	T	60.05	73.34	172.37	118.09	-	51	K	56.27	33.13	176.25	123.46	8.30
23	P	63.67	32.17	176.95	139.50	-	52	K	54.40	32.33	174.49	124.98	8.41
24	K	56.41	32.61	176.78	122.63	8.50	53	P	63.19	32.14	176.45	137.34	-
25	K	56.38	33.02	176.18	123.28	8.35	54	A	52.46	19.15	177.28	124.64	8.34
26	A	52.53	19.16	177.22	125.69	8.29	55	A	52.44	19.29	177.43	123.94	8.27
27	A	52.50	19.10	177.47	124.11	8.25	56	T	6-	73.41	172.40	117.48	8.54
28	K	54.20	32.30	174.54	122.22	8.28	57	P	63.55	32.25	177.03	139.79	-
29	P	63.25	32.06	176.85	137.22	-	58	K	56.39	32.29	176.82	122.64	8.53

59	K	56.23	33.00	176.21	122.98	8.35	**80**	K	56.65	33.04	176.71	121.39	8.35
60	A	52.50	19.09	177.54	125.61	8.28	**81**	T	62.30	70.15	174.24	116.82	8.20
61	K	56.42	33.03	176.37	121.47	8.28	**82**	V	62.20	32.64	175.70	123.75	8.24
62	K	54.39	32.52	174.43	124.32	8.33	**83**	K	54.37	32.42	174.36	127.28	8.42
63	P	63.10	32.13	176.63	137.17	-	**84**	P	63.26	31.95	176.86	137.12	-
64	K	56.73	32.97	176.49	122.31	8.37	**85**	K	56.39	32.95	176.51	122.12	8.38
65	V	62.51	32.78	175.83	122.97	8.19	**86**	A	52.56	19.17	177.81	125.81	8.32
66	V	62.44	33.02	175.78	125.82	8.30	**87**	K	56.97	33.57	176.86	121.30	8.37
67	K	56.31	33.31	176.18	126.49	8.41	**88**	S	58.39	64.13	174.75	117.18	8.35
68	V	62.46	32.87	175.77	122.50	8.16	**89**	S	-	-	174.38	118.16	8.38
69	K	54.45	32.33	174.39	127.24	8.40	**90**	A	53.26	19.03	177.83	126.12	8.26
70	P	63.04	32.13	176.82	137.02	-	**91**	K	-	-	176.77	120.75	8.22
71	V	62.55	33.00	176.24	121.04	8.22	**92**	R	56.45	33.24	176.10	122.58	8.33
72	K	56.53	32.92	176.02	125.65	8.36	**93**	A	52.42	19.06	177.78	125.67	8.35
73	A	52.71	19.12	177.57	126.32	8.35	**94**	S	-	-	174.71	115.66	8.27
74	S	58.51	64.28	174.12	116.12	8.30	**95**	K	56.63	32.43	176.63	123.63	8.35
75	K	54.38	32.48	174.43	124.31	8.31	**96**	K	63.33	31.97	176.81	122.61	8.28
76	P	63.13	32.16	176.81	137.32	-	**97**	K	56.43	32.74	176.63	121.86	8.40
77	K	56.52	32.75	176.68	122.36	8.42	**98**	R	55.87	31.19	176.19	122.91	8.30
78	K	56.36	33.18	176.13	123.18	8.31	**99**	S	58.67	64.18	173.58	118.27	8.41
79	A	52.40	19.18	177.60	126.28	8.34	**100**	H	-	-	-	125.39	-

Appendix 2. The sample was prepared at 1mM protein concentration in H_2O/D_2O 9:1 v/v at 302.1K and pH 5.5 conditions. (-) Not observed due to fast HN <-> H_2O fast exchange or signal overlap.

III.   APPENDIX

Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and Virus Strains		
E. Coli: BL21(DE3) chemically competent cells	Stratagene	200131
E. Coli: XL10-Gold chemically competent cells	Stratagene	200314
E. Coli: BL21-DE3 RF10 auxotrofic strain	Addgene	
E. Coli: BL21-DE3 RF6 auxotrofic strain	Addgene	
Biological Samples		
Protein Pub RRM 123	This thesis	N/A
Protein Pub RRM 12	(Santiveri et al., 2011)	N/A
Protein Pub RRM 3	(Santiveri et al., 2011)	N/A
Protein Pab RRM 1	This thesis	N/A
Protein Pab RRM 2	This thesis	N/A
Protein Pab RRM 12	This thesis	N/A
Protein eIF4G1$_{1-249}$	This thesis	N/A
Protein eIF4G1$_{1-184}$	This thesis	N/A
Protein eIF4G1$_{1-249}$ (ΔBox1)	This thesis	N/A
Protein eIF4G1$_{1-184}$ (ΔBox1)	This thesis	N/A
Protein eIF4G1$_{1-184}$ (ΔBox2)	This thesis	N/A
Protein eIF4G1$_{188-249}$	This thesis	N/A
Protein eIF4G$_{35-51}$-Pub RRM3	This thesis	N/A
Chemicals, Peptides, and Recombinant Proteins		
Pentaethylene glycol monododecyl ether, $C_{12}E_5$	Sigma-Aldrich	CAS:3055-95-6
Hexaethylene glycol monododecyl ether, $C_{12}E_6$	Sigma-Aldrich	CAS:3055-96-7
Filamentous phage Pf1	ASLA biotech	P-50-RNA
Ficoll PM70	GE Healthcare	Product:17031005

4-(2-Iodoacetamido)-TEMPO	Sigma-Aldrich	CAS:25713-24-0
Deuterium Oxide D$_2$O	Euriso-top	CAS:7789-20-0
Lysozyme from chicken egg white	Sigma-Aldrich	CAS:12650-88-3
Protease Inhibitors cOmplete ultra tablets	Roche	REF. 05892791001
Isopropylthio-ß-galactoside, IPTG	Generon	Cat# GF101-01
1,4-Dithiothreitol, DTT	Sigma-Aldrich	CAS:3483-12-3
2-Mercaptoethanol	Sigma-Aldrich	CAS:60-24-2
Imidazole, 99%	ACROS Organics^{TM}	CAS:288-32-4
Sodium chloride	Fisher Chemical	CAS:7647-14-5
Potassium Phosphate Dibasic	Fisher BioReagents	CAS:7758-11-4
Potassium Phosphate Monobasic	ACROS Organics^{TM}	CAS:7778-77-0
Yeast Extract	CONDA pronadisa	Cat#1702-00
Tryptone	CONDA pronadisa	Cat#1612-00
Ammonium chloride (15N, 99%)	Cambridge Isotope Lab.	CAS:39466-62-1
D-Glucose (U-13C6, 99%)	Cambridge Isotope Lab.	CAS:110187-42-3
MOPS	Sigma-Aldrich	CAS:1132-61-2
L-cysteine	Fisher Chemical	CAS:52-90-04
L-tyrosine	Merck	CAS:60-18-4
L-Alanine	Merck	CAS:56-41-7
DL-Tryptophan	Merck	CAS:87-32-1
L-Arginine hydrochloric	Carlo Erba	CAS:1119-34-2
L-Leucine ISO	Carlo Erba	CAS:73-32-5
Acid Aspartic	Carlo Erba	CAS:56-84-8
L-Lysine monochroic	Carlo Erba	CAS:657-27-2
L-Leucine	Carlo Erba	CAS:61-90-5
DL-Methionine	Carlo Erba	Cod:463122
DL-Serine	Carlo Erba	Cod:477908
L-Proline	Carlo Erba	Cod:474707
L-Asparagine	Carlo Erba	Cod:424542

L-Histidine monochroic	Carlo Erba	Cod:456958
Glycine	Sigma-Aldrich	CAS:56-40-6
DL-Threonine	Sigma-Aldrich	CAS:80-68-2
L-Glutamine	Sigma-Aldrich	CAS:56-85-9
L-Histidine	Sigma-Aldrich	CAS:71-00-1
L-Glutamic acid monosodium salt hydrate	Sigma-Aldrich	CAS:6106-04-3
Tricine	Sigma-Aldrich	CAS:5704-04-1
Magnesium sulfate	Fluka	CAS:17830-18-1
Calcium chloride	Panreac	Cod:141219
Ferrous chloride	Probys	Cod:53198
Cobalt chloride	Fluka	CAS:7791-13-1
Cooper chloride	UCB	Cod:1277
Manganese chloride	UCB	Cod:1436
Zinc chloride	Fluka	CAS:7646-85-7
Sodium molybdate dehydrate	Sigma-Aldrich	CAS:10102-40-6
Thiamine hydrochloride	Sigma-Aldrich	CAS:67-03-8
d-Biotine	Sigma-Aldrich	CAS:58-85-5
Choline chloride	Sigma-Aldrich	CAS:67-48-1
Folic Acid	Sigma-Aldrich	CAS:59-30-3
Niacinamide	Sigma-Aldrich	CAS:98-92-0
d-pantothenic	Sigma-Aldrich	CAS:137-08-6
Pyridoxal	Sigma-Aldrich	CAS:65-22-5
Riboflavin	Sigma-Aldrich	CAS:83-88-5
Tris	Fisher	CAS:7786-1
Kanamycin sulfate	Calbiochem	Cat#420411
Boric acid	Sigma-Aldrich	CAS:10043-35-3
PIPES	Sigma-Aldrich	CAS:5625-37-6
Potassium chloride	Fisher	CAS:7447-40-7
Dimethyl sulfoxide, DMSO	Carlo Erba	Cod:445103

Tris hydroxymethyl aminomethane hydrochloride	ACROS Organics™	CAS:1185-53-1
Sodium dodecyl sulfate, SDS	Silga-Aldrich	CAS:151-21-3
Acrylamide	Scharlau	CAS:79-06-1
N,N,N',N'-tetra methyl-ethylenediamine, TEMED	Sigma-Aldrich	CAS:110-18-9
Agarose LM	Pronadisa	Cat#8051
Alexa Fluor 488	Invitrogen	Cat#A-11001
Alexa Fluor 647	Invitrogen	Cat#A-21235
Agarose EEO	Pronadisa	Cat#8022
TEV protease	Homemade	N/A

Critical Commercial Assays

Taq Master Kit	Jean Bioscience	Cat#PCR-101L
Kit PCR Purification	Jean Bioscience	Cat#PP-2015
New Builder Hifi DNA Assembly Master Mix	BioLabs	Cat#D2621L
Bug Buster Master Mix	Millipore	Cat#71456
QuikChange Lightning Kit	Agilent genomics	Ref.210515
DNA polymerase KOD	Novagen	Ref. 71085-3
DNA polymerase Pfu	Promega	Ref. M774A
Fast n-Easy Plasmid Miniprep Kit	Jean Bioscience	Cat#PP-204S

Deposited Data

BMRB	Pab $_{1-213}$	
BMRB	eIF4G1 $_{1-250}$	
BMRB	Quimera	

Experimental Models: Organisms/Strains

Saccharomyces cerevisiae	Novagen	Ref. 69240-3

Oligonucleotides

PUB1_R12_GGGS_Rv:GTGGTGGGATCCGCCTCCATCACGCTTAGCAGCCCAGGTG	(Santiveri et al., 2011)	IDT
PUB1_R3_GGGS_Rv:GTGGTGGGATCCGCCTCCTCTTTACTTACCCCAACCGGTTCTC	(Santiveri et al., 2011)	IDT
PUB_31_FW:CCAGGATCCGCAGATCCTTCTTCTGAACAGAGCGTCGCTGTCGAAGGC	(Santiveri et al., 2011)	PROLIGO

PUB_242_RV:CGCGCTCGAGTTAATCACGCTTAGCAGCCCAGTTGATTCTTAGCGGTC	(Santiveri et al., 2011)	PROLIGO
PUB1_315_FW:CGCGGATCCCAGACCATTGGTTTACCTCCTCAAGTAAATCCTCAAGC	(Santiveri et al., 2011)	SIGMA
PUB1_414_RV:GCGCTCGAGTTATCTTTCCTTACCCCAACCGGTTCTCAACTTTCTGCG	(Santiveri et al., 2011)	SIGMA
PUB_31_FW:CCAGGATCCGCAGATCCTTCTTCTGAACAGAGCGTCGCTGTCGAAGGC	(Santiveri et al., 2011)	SIGMA
PUB1_414_RV:GCGCTCGAGTTATCTTTCCTTACCCCAACCGGTTCTCAAGTTTCTGCC	(Santiveri et al., 2011)	SIGMA
PUB1_75_FW:CTGTATTTCCAGGGATCCAGAGTTTTATATGTTGGTAACCTAG	This thesis	MACROGEN
PUB1_75_RV:TTACCAACATATAAAACTCTGGATCCCTGGAAATACAGG	This thesis	MACROGEN
PUB1_161_FW:CTGTATTTCCAGGGATCCACATTTAACTTGTTTGTCGGTGATTG	This thesis	MACROGEN
PUB1_161_RV:TCACCGACAAACAAGTTAAATGTGGATCCCTGGAAATACAGG	This thesis	MACROGEN
PAB1_1_FW:GCGCGGATCCATGGCTGATATTACTGATAAGACAGCTGAACAATTGG	This thesis	IDT
PAB1_end_RV:CGCGCTCGAGTAGGGAAGTAGGTGATTACATAGAGC	This thesis	IDT
PAB1_208STBAM_FW:CCAGAAAGGAACGTTAAGGATCCTTGGAAGAGACTAAGGC	This thesis	IDT
PAB1_208STBAM_RV:GCCTTAGTCTCTTCCAAGGATCCTTAACGTTCCTTTCTGG	This thesis	IDT
PAB1_487STBAM_FW:GGTTTACGGCGTCCCACCATAAGGATCCTTCCCAAGAAATGCC	This thesis	IDT
PAB1_487STBAM_RV:GGCATTTCTTGGGAAGGATCCTTATGGTGGGACGCCGTAAACC	This thesis	IDT
PAB1_405STBAM_FW:GAAAAGACGTAAGACGTTAAGGATCCGCTCAACAAATCCAAGC	This thesis	IDT
PAB1_405STBAM_RV:GCTTGGATTTGTTGAGCGGATCCTTAACGTCTTACGTCTTTTC	This thesis	IDT
PAB1_L261F_FW:AATTGAAGGGTTTCGGGTTTGT	This thesis	IDT
PAB1_L261F_RV:ACAAACCCGAAACCCTTCAATT	This thesis	IDT
PAB1_123STBAM_FW:CCATCATTGAGAAAGTAAGGATCCGGTAACATCTTTATCAAGAAC	This thesis	IDT
PAB1_123STBAM_RV:GTTCTTGATAAAGATGTTACCGGATCCTTACTTTCTCAATGATGG	This thesis	IDT
PAB1_R2_GGGS2_RV:AGTGGATCCACCTCCACGTTCCTTTCTGGACAAGTG	This thesis	IDT
PAB_36_FW:CTGTTGAAAACTCTTCTGCATCATTATATGTTGG	This thesis	IDT
PAB1_R2_GGGS_RV:GTGGTGGGATCCGCCTCCTCTTTCCTTACCCCAACCGGTTCTC	This thesis	IDT

PAB1_36_Bgl_FW:CGCAGATCTGTTGAAAACTCTTCTGCATCATTATAT	This thesis	IDT
PAB1_218_ST_REV:GCGCTCGAGTTAATGTGCCTTAGTCTCTTCC	This thesis	IDT
PAB1_113_FW:CGGAGATCTTGGTCTCAACGTGACCCATCATTGAGA	This thesis	IDT
PAB1_113ST_FW:ACTCTCAATTGGAATAAACTAAGGCACATTACAC	This thesis	MACROGEN
PAB1_113ST_RV:GTAATGTGCCTTAGTTTATTCCAATTGAGAGTC	This thesis	MACROGEN
PAB1_35_FW:CCTGTATTTCCAGGGATCCAACTCTTCTGCATCATTATATG	This thesis	MACROGEN
PAB1_35_RV:CATATAATGATGCAGAAGAGTTGGATCCCTGGAAATACAGG	This thesis	MACROGEN
PAB1_180KE_ER181_FW:GAAGAAGGTGCTGCCGAGAAAGCTATTGATGCTTTG	This thesis	MACROGEN
PAB1_180KE_ER181_RV:CAAAGCATCAATAGCTTTCTCGGCAGCACCTTCTTC	This thesis	MACROGEN
PAB1_202ST_FW:GCTCCTCACTAATCCAGAAAGGAACG	This thesis	MACROGEN
PAB1_202ST_RV:CGTTCCTTTCTGGATTAGTGAGGAGCAAC	This thesis	MACROGEN
PAB1_30_FW:CTGTATTTCCAGGGATCCCAATCTGTTGAAAACTCTTC	This thesis	MACROGEN
PAB1_30_RV:AGTTTTCAACAGATTGGGATCCCTGGAAATACAGG	This thesis	MACROGEN
TIF4631_1_FW:CGCGGATCCATGACAGCAGAAACTGCTCACCCGACACAATCTGC	(Santiveri et al., 2011)	IDT
TIF4631_394_ST_RV:CGCGCTCGAGTTAATCAGTTGTAGTTTCGATTTCAGCTTCAAGTCC	(Santiveri et al., 2011)	IDT
TIF4631_82_stop_RV:CGCGCTCAGATTATCTGAAACTACCGCCACC	(Santiveri et al., 2011)	IDT
TIF4631_83_FW:CGCGGATCCGGTGGACACATGGGAGCCAACAGC	(Santiveri et al., 2011)	IDT
TIF4631_187_FW:CGCGGATCCACTAATGACTCTAAGGCCAGTTCTG	(Santiveri et al., 2011)	IDT
TIF4631_186_stop_RV:GCGCTCGAGTTAAGGGGTAGGAGTTGGAGTAGAAG	(Santiveri et al., 2011)	IDT
TIF4631_82_STBAM_FW:GGCGGTAGTTTCAGATAAGGATCCATGGGAGCCAACAGC	(Santiveri et al., 2011)	IDT
TIF4631_82_STBAM_RV:GCTGTTGGCTCCCATGGATCCTTATCTGAAACTACCGCC	(Santiveri et al., 2011)	IDT
TIF4631_188STBAM_FW:CTTCTACTCCAACTCCTTAAGGATCCACTAATGACTCTAAGG	(Santiveri et al., 2011)	IDT
TIF4631_188STBAM_RV:CCTTAGAGTCATTAGTGGATCCTTAAGGAGTTGGAGTAGAAG	(Santiveri et al., 2011)	IDT

TIF4631_229STBAM_FW:GAGAAAGGAGCAACTTTA AGGATCCAGTGGCAACAATAATATTCC	(Santiveri et al., 2011)	IDT
TIF4631_229STBAM_RV:GGAATATTATTGTTGCCAC TGGATCCTTAAAGTTGCTCCTTTCTC	(Santiveri et al., 2011)	IDT
TIF4631_114STBAM_FW:ATGGCGGCCTAAGGATCT GCCCCTGCTAATCC	This thesis	IDT
TIF4631_W95A_FW:GCTCAAACGTGCCAGCGACTG GTTACTATAATAAC	This thesis	IDT
TIF4631_W95A_RV:GTTATTATAGTAACCAGTCGCT GGCACGTTTGAGC	This thesis	IDT
TIF4631_F98_99A_FW:GTGCCATGGACTGGTGCCG CTAATAACTACCCCG	This thesis	IDT
TIF4631_F98_99A_RV:CGGGGTAGTTATTAGCGGCA CCAGTCCATGGCAC	This thesis	IDT
TIF4631_F105_106A_FW:CTATAATAACTACCCCGTT GCCGCCCAGCCCCAGCAA	This thesis	IDT
TIF4631_F105_106_A_RV:TTGCTGGGGCTGGGCGG CAACGGGGTAGTTATTATAG	This thesis	IDT
TIF4631_114STBAM_RV:GGATTAGCAGGGGCAGAT CCTTAGGCCGCCAT	This thesis	IDT
TIF4631_F105_106A_2_FW:CTATAATAACTACCCCG TTGCCGCCCAGCCCCAGCAA	This thesis	IDT
TIF4631_F105_106A_2_RV:TTGCTGGGGCTGGGCG GCAACGGGGTAGTTATTATAG	This thesis	IDT
TIF4631_DBOX1_FW:CCAACAGCTCAAACGTGGCGG	This thesis	IDT
TIF4631_DBOX1_RV:CCGCCACGTTTGAGCTGTTGG	This thesis	IDT
TIF4631_DBOX2_FW:CCAATTCCTGTCGAAGAGAAG AAAGCCAAGCTACAGTCTCAGG	This thesis	IDT
TIF4631_DBOX2_RV:CCTGAGACTGTAGCTTGGCTTT CTTCTCTTCGACAGGAATTGG	This thesis	IDT
TIF4631_82_BAM_RV:CGCGGATCCTCTGAAACTAC CGCCACC	This thesis	IDT
TIF4631_41STBAM_FW:CACCAACTACAACTAAGGA TCCAACTATACCCAAAAG	This thesis	IDT
TIF4631_160STBAM_FW:CAGTCTCAGGAAAGATAA ACT	This thesis	IDT
TIF4631_160STBAM_RV:GACTCTGGTTGCGGGGAT CCA	This thesis	IDT
TIF4631_S90C_FW:ATGGGAGCCAACTGCTCAAACG TGCCA	This thesis	IDT
TIF4631_S90C_RV:TGGCACGTTTGAGCAGTTGGCTC CCAT	This thesis	IDT
TIF4631_S30C_FW:CTTCAAAGCAGGAATGTGCTGC TCTGAAAC	This thesis	IDT
TIF4631_S30C_RV:GTTTCAGAGCAGCACATTCCTGC TTTGAAG	This thesis	IDT

TIF4631_S128C_FW:GTCGAAGAGAAGTGTCCTGTT CCAACTAAG	This thesis	IDT
TIF4631_S128C_RV:CTTAGTTGGAACAGGACACTTC TCTTCGAC	This thesis	IDT
TIF4631_S163C_FW:AGATCTACTGTGTGTCCGCAAC CAGAG	This thesis	IDT
TIF4631_S163C_RV:CTCTGGTTGCGGACACACAGTA GATCT	This thesis	IDT
TIF4631_119ST_RV:GCGCTCGAGTTAAGGGGCACTA CCAGCGGCCGC	This thesis	IDT
TIF4631_305ST_RV:GCGCTCGAGTTACTTCAAACGT TCAGCAAAGGTTAAC	This thesis	IDT
TIF4631_348ST_RV:GCGCTCGAGTTATTTAACCTGTT CACTGGGAGGC	This thesis	IDT
TIF4631_250ST_FW:GTGGAAGAGAAGTAATCGGAC AAACCTG	This thesis	MACROGEN
TIF4631_250ST_RV:GGTTTGTCCGATTACTTCTCTTC CACATTTTCTGG	This thesis	MACROGEN
TIF4631_274ST_FW:GAGCCAGAAGTTAAGTAAGAA ACTCCAGCTGAAG	This thesis	MACROGEN
TIF4631_274ST_RV:CAGCTGGAGTTTCTTACTTAACT TCTGGCTCAGC	This thesis	MACROGEN
TIF4631_G35S_FW:CTCAGCAACAACGTAGCTACAC CAACTACAAC	This thesis	MACROGEN
TIF4631_G35S_RV:GTAGTTGGTGTAGCTACGTTGTT GCTGAGATTCC	This thesis	MACROGEN
TIF4631_G65S_FW:GAGGTGGTAAATTTAGTCCAAA CAGATATAACAACC	This thesis	MACROGEN
TIF4631_G65S_RV:GTTATATCTGTTTGGACTAAATTT ACCACCTCTTTGC	This thesis	MACROGEN
TIF4631_G35C_FW:CTCAGCAACAACGTTGCTACAC CAACTACAA	This thesis	MACROGEN
TIF4631_G35C_RV:GTAGTTGGTGTAGCAACGTTGTT GCTGAGATTCC	This thesis	MACROGEN
TIF4631_G65C_FW:GGTGGTAAATTTTGTCCAAACA GATATAACAACC	This thesis	MACROGEN
TIF4631_G65C_RV:GTTATATCTGTTTGGACAAAATT TACCACCTCTTTGC	This thesis	MACROGEN
TIF4631_115_FW:CTGTATTTCCAGGGATCCGCCCCT GCTAATCCAATTCC	This thesis	MACROGEN
TIF4631_115_RV:GGATTAGCAGGGGCGGATCCCTG GAAATACAGGTTTTC	This thesis	MACROGEN
TIF4631_Q109C_FW:TACCAGCCCTGCCAAATGGCG GCC	This thesis	MACROGEN
TIF4631_Q109C_RV:CCGCCATTTGGCAGGGCTGGT AGTAAACG	This thesis	MACROGEN

TIF4631_G230C_FW:GGAGCAACTTGAATGTTCTAG TGGCAAC	This thesis	MACROGEN
TIF4631_G230C_RV:GCCACTAGAACATTCAAGTTGC TCC	This thesis	MACROGEN
TIF4631_S200C_FW:CTGAAGAAAATATATGTGAAG CTGAAAAGACAAG	This thesis	MACROGEN
TIF4631_S200C_RV:CTTTTCAGCTTCACATATATTTT CTTCAGAACTGG	This thesis	MACROGEN
TIF4631_114STBAM_FW2:ATGGCGGCCTAAGGATC CGCCCCTGCTAATCC	This thesis	MACROGEN
TIF4631_114STBAM_RV2:ATTAGCAGGGGCGGATC CTTAGGCCGCCATTTGC	This thesis	MACROGEN
TIF4631_235ST_FW:GGTTCTAGTGGCAACTAAAATA TTCCAATGAAGACTACC	This thesis	MACROGEN
TIF4631_235ST_RV:CTTCATTGGAATATTTTAGTTGC CACTAGAACCTTCAAG	This thesis	MACROGEN
TIF4631_199_FW:ATTTCCAGGGATCCATATCTGAAG CTGAAAAGACAAGAAG	This thesis	MACROGEN

Recombinant DNA

Plasmid: pET28_txaHTEV_PUB1_R12_GGGS_R3	This thesis	N/A
Plasmid: pET28_txaHTEV_PUB1_R3	This thesis	N/A
Plasmid: pET28_txaHTEV_PUB1_R12	This thesis	N/A
Plasmid: pET28_txaHTEV_PAB1_R1	This thesis	N/A
Plasmid: pET28_txaHTEV_PAB1_R2	This thesis	N/A
Plasmid: pET28_txaHTEV_PAB1_R12	This thesis	N/A
Plasmid: pET28_txaHTEV_TIF4631	This thesis	N/A

Software and Algorithms

TopSpin 2.1	Bruker	https://www.bruker.com
TopSpin 3.5pl7	Bruker	https://www.bruker.com
nmrPipe	IBBS	https://www.ibbr.umd.edu/nmrpipe/index.html
ASTRA SEC-software version 5.3.4	Wyatt Technology	https://www.wyatt.com
Dynamics	Wyatt Technology	https://www.wyatt.com

CcpNmr Analysis	Skinner et al., 2016, Vranken et al., 2005	https://www.ccpn.ac.uk
Flexible Meccano	Ozenne et al., 2012	V1.1
Pymol	Schödinger	https://pymol.org/
ApE	Plasmid Editor by M. Wayne Davis	jorgensen.biology.utah.edu/wayned/ape/
PRIMUS	Konarev et al., 2003	N/A
GREMLIN	Kamisetty et al., 2013	N/A
CRYSOL	Svergun et al., 1995	N/A
Other		
Vivaspin 20, 5kDa MWCO concentrators	Sigma Aldrich	GE28-9323-59
Vivaspin Turbo 15, 10kDa MWCO concentrators	Sartorius	Order Nº VS15T01
Vivaspin 20, 30kDa MWCO concentrators	Sartorius	
Vivaspin 6, 10kDa MWCO concentrators	Sartorius	Order Nº VS0602
Hitrap™ IMAC FF	GE Healthcare	Cod:17-0921-04
Hitrap™ Q HP	GE Healthcare	Cod:17-1154-01
Hitrap™ SP HP	GE Healthcare	Cod:17-1152-01
Bio-Scale™ bio-gel P-6 Desalting Cartridge	BIO-RAD	Cat#732-5312
Nap™ -5 Columns Sephadex™ G-25 DNA Grade	GE Healthcare	Cod:17-0853-01
Superdex™ 200 10/300 GL	GE Healthcare	Cod:17-5157-01
Differential Refractive Index Detector	Shimadzu	RID-20A
MALS detector	Wyatt Techonology Corp	Dawn Heleos II

Appendix 3. This table highlights the genetically modified strains, reagents, software and source data essential to reproduce results presented in the Chapter 5, 6 and 7.

IV. APPENDIX

NMR spectral assignment and interactions of the N-terminal Instrinsically unstructured domain of eIF4G

Chaves–Arquero, Belén,[a] Martínez–Lumbreras, Santiago[a,b] and Pérez–Cañadillas, Jose Manuel[a]

[a] Departamento de Química-Física Biológica. Instituto de Química-Física "Rocasolano". CSIC. C/Serrano 119, 28006 Madrid. Spain.
[b] Present address: Department of Chemistry. School of Biomedical Science. King's College London. 7 Trinity Street. London. SE1 1DB. United Kingdom

Introduction

Eukaryotes have a large percentage of proteins that are predicted to be fully intrinsically disordered or to have large disordered regions alternating with folded domains. These intrinsically disordered proteins (IDPs) play key roles in a range of biological processes, including celular signaling or gene expression regulation. A protein of this type is the eukaryotic translation initiation factor eIF4G (also known as Tif4631 in yeast), which plays a positive role in translation regulation by enhancing ribosome recruitment to the mRNA, but also a negative role by being an essential components of stress granules.

Tif4631 N-terminal domain (1–402) is an IDP that contains several Molecular Recognition Fragments or MoRFs, some of them involved recognition of RNA binding proteins like Pub1 and Pab1 and some others involved in self-recognition and aggregation. Previous studies in our group have proposed an interaction model for Tif4631 (Figure 1). We also have previously assigned the construct Tif4631 (1-187). Here we present the NMR study of a longer construct, Tif4631 (1-249), containing all the MoRFs of the proposed model (Figure 1).

To achieve this very challenging study we have combined standard NMR approaches with others like residue-specific isotope labelling and paramagnetic relaxation enhancement (PRE).
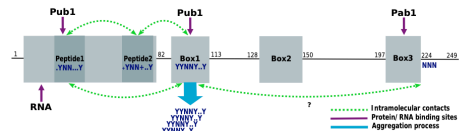
Figure 1. Proposed network of interactions be among Pab1, Pub1 and Tif4631 (N-terminal domain)

Assignment

NMR spectral assignment of Tif4631 (1-249)

To facilitate NMR assignment, we use an amino acid-selective isotope labeling, which is a powerful approach in structural studies of proteins[2]. Escherichia Coli auxotrophic strains RF6 and RF10 has been used for production of Pro or Lys reverse-labelled Tif4631 (1-249) samples (^{15}N/^{13}C labelled at residues different of Pro or Lys). The isotope-edited [^{1}H,^{15}N]–HSQC greatly simplifies the complexity of the spectra (all of them have been assigned already) while still map different sites along the 1-249 sequence. This approach can be used to simplify interaction studies by chemical shift mapping or PRE.

A) [^{1}H,^{15}N]–HSQC spectrum of Tif4631 (1-249). B) [^{1}H,^{15}N]–HSQC spectrum with isotopic edition of NH signals following Lys residues (^{12}C/^{14}N unlabelled). C) Equivalent isotope-edited spectrum for the reverse labelled sample of Pro.
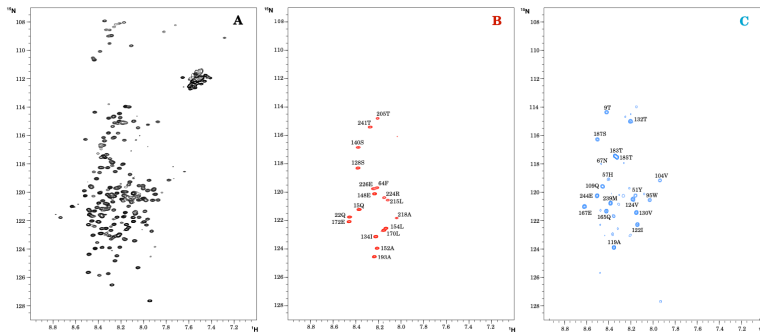
Interactions

Mapping Tif4631 intramolecular interactions by paramagnetic Relaxation Enhancement (PRE)

We introduced spin labels at strategic positions within Tif4631 (1-249) via Gly/Cys or Ser/Cys mutations. These single cysteine mutants (wild-type sequence is devoid of Cys) were alkylated with 4-(2-Iodoacetamido)-TEMPO as described[3], to introduce the unpair electron that allows to map distances by Paramagnetic Relaxation Enhancement in the 10 to 20 Å range.
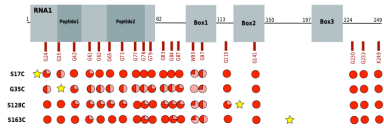

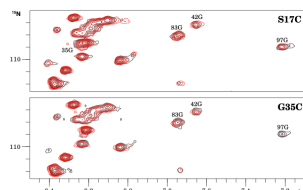Figure 2. Extent of the PRE on the Gly signals for several Tif4631 mutants (S17C, G35C, S128C, S163C)

Figure 3. The [^{1}H,^{15}N]–HSQC spectrum of the TEMPO adduct (red) was compared with spectra of the same sample after reduction of the nitroxy group with sodium ascorbate (grey).

The preliminary analysis of the PRE data for mutants S17C and G35C is compatible with the proposed Pep1-Box1 intramolecular contact (Figure 1).

Pub1- Tif4631 interaction

Pub1 is a 435–residue RNA binding protein (three RRM) involved in the formation of stress granules. Previous studies in our lab[4] have mapped the interaction between Tif4631 and Pub1 RRM3. Peptides from regions Pep1 and Box1 reproduce the pattern of chemical shift changes observed for Tif4631 (1-402). Interestingly our PRE data show that these two regions interacts in Tif4631 free-state.

When monitoring the Tif4631-Pub1 interaction on the Pub1 [^{1}H,^{15}N]–HSQC spectra two different scenarios occurred:

Tif4631 1-82, 1-249, 1-305, 1-348 and 1-402 cause chemical shift perturbations on fast exchange (compatible with weak binding); however, Tif4631 1-228 and 1-184 cause a global collapse of the signals (soluble aggregates). Mutants behaviour demonstrate that aggregation is dependent on the Box1 element. A global analysis of the data suggest that Tif4631 could have an intrinsic propensity to aggregate that can be modulated by Pub1 and perhaps other RNA binding proteins (i.e. Pab1). This physical-chemical feature might represent an important event during the phase separation stages that lead to rapid formation of stress granules in vivo.
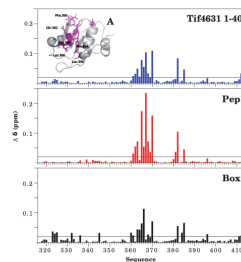

Figure 4. Chemical shift mapping of Tif4631p (proteins and peptides) on the spectra of Pub1p RRM3. A) NMR chemical shift mapping (backbone amide groups) of the titration of Pub1p RRM3 with Tif4631p (1-402).
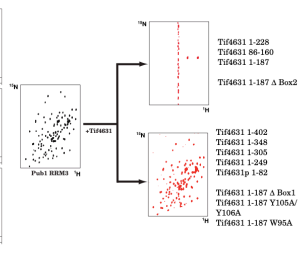
Figure 5. The figure shows different [^{1}H,^{15}N]–HSQC spectra of labelled Pub1 RRM3 construct titrated with several Tif4631 constructs (red), overlapped with the free spectrum (grey)

References

1. Santiago, Martínez-Lumbreras.The role of Gbp2p, Nab2p and Pub1p along the mRNA cycle: structural and molecular recognition sudies by NMR and other biophysiscal techniques. (2013) PhD Thesis

2. Myat T. Lin, Risako Fukazawa, Yoshihary Miyajima-Nakano, Shinichi Matsushita, Sylvia K. Choi, Toshio Iwasaki, Robert B. Gennis.. Escherichia coli Auxotroph Host Strains for Amino Acid-Selective Isotope Labeling of Recombinant Proteins. (2015) Methods in Enzymology, Volume 565.

3. Joel R. Gillespie and David Shortle. Characterization of Long-range Structure in the Denatured State of Staphylococcal Nuclease I. Paramagnetic Relaxation Enhancement by Nitroxide Spin Labels. (1997) J. Mol. Biol. 268, 158-169.

4. Santiveri CM, Mirassou Y, Rico-Lastres P, Martínez-Lumbreras S, Pérez-Cañadillas JM. Pub1p C-terminal RRM domain interacts with Tif4631p through a conserved region neighbouring the Pab1p binding site. PLoS One. 2011; 6(9):e24481

Loading poly(U) and poly(A) RNA binding proteins on the N-terminal intrinsically unstructured domain of eIF4G

Chaves–Arquero, Belén,[a] Martínez–Lumbreras, Santiago[a,b] and Pérez–Cañadillas, Jose Manuel[a]

[a] Departamento de Química-Física Biológica. Instituto de Química-Física "Rocasolano". CSIC. C/Serrano 119, 28006 Madrid. Spain.
[b] Present address: Department of Chemistry. School of Biomedical Science. King's College London. 7 Trinity Street. London. SE1 1DB. United Kingdom

Introduction

Many eukaryotic proteins are intrinsically disordered (IDPs) or have large disordered regions that play key biological roles processes like celular signaling or gene expression regulation. The N-terminal domain (1-402) of eIF4G (known as Tif4631 in yeast) is an IDP that regulate translation in positive (ribosome recruitment) and negative (formation of stress granules) ways. Tif4631 N-terminal domain contain sequences for recognition of poly(U) and poly(A) RNA binding proteins (Pub1 and Pab1), self-recognition and aggregation (Figure 1). Here we present the NMR study of Tif4631 (1-249) focusing on intra and intermolecular interactions.

Figure 1. Proposed the Molecular Recognition Fragments or MoRFs of Tif4631 (1-249)

To study intramolecular interactions and conformational properties of Tif4631 (1-249) we introduced spin labels via Gly/Cys or Ser/Cys mutations that were alkylated with 4-(2-Iodoacetamido)-TEMPO as described[3], to introduce the unpaired electron that allows mapping distances by Paramagnetic Relaxation Enhancement in the 10 to 20 Å range. Here we show a preliminary analysis on the Gly signals. A more detailed analysis including the rest of resonances is currently being done.
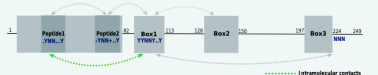
The comparison of the PRE data for mutants S17C and G35C suggest the presence of an intramolecular interaction between the Pep1 and BOX1 regions (Figure 2 and scheme).

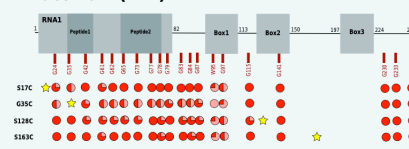### Mapping intramolecular interactions in Tif4631 (1-249) in by Paramagnetic Relaxation Enhancement (PRE)

Figure 2. Extent of the PRE on the Gly signals for several Tif4631 mutants (S17C, G35C, S128C, S163C)
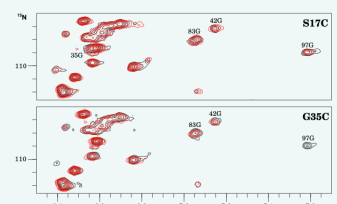

Figure 3. The [¹H,¹⁵N]–HSQC spectrum of the TEMPO adduct (red) was compared with spectra of the same sample after reduction of the nitroxy group with sodium ascorbate (grey)

Interactions between Tif4631 (1-249) and RNA binding proteins Pub1 and Pab1

Pub1-Tif4631 interaction

Previous studies in our lab[4] have mapped the interaction between Tif4631 to the third RRM of Pub1. Further studies show that the interaction occurs with regions Pep1 and Box1 (Figure 4). These two peptides and several constructs of Tif4631 induce chemical shift perturbations on the Pub1 RRM3 spectra in fast exchange regime. But surprisingly, Tif4631 1-228 and 1-184 constructs lead to extreme signal broadening. Further mutagenesis studies show that Box1 is the element responsible for this behaviour. We reasoning that this global collapse of the signals is due to formation of oligomers or soluble aggregates, a physical-chemical process that might perhaps being important during the earlier stages of stress granules nucleation.
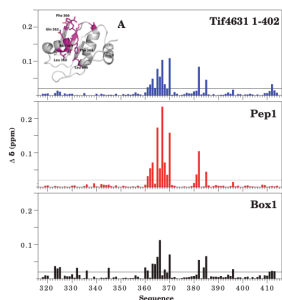

Figure 4. Chemical shift mapping of Tif4631 (proteins and peptides) on the spectra of Pub1 RRM3. A) NMR chemical shift mapping (backbone amide groups) of the titration of Pub1 RRM3 with Tif4631 (1-402)
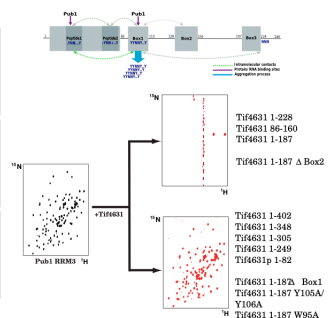
Figure 5. The figure shows different [¹H,¹⁵N]-HSQC spectra of labelled Pub1 RRM3 construct titrated with several Tif4631 constructs (red), overlapped with the free spectrum (grey)

Pab1-Tif4631 interaction

To study Pab1/Tif4631 interaction we used amino acid-selective isotope unlabeling[6] that allows simplifying the ¹H-¹⁵N HSQC spectra by using isotopically discriminated NMR spectroscopy[7]. Binding cause severe line broadening of the Tif4631 (1-249) signals from BOX1, BOX2 and BOX3 either due to chemical exchange or to formation of large molecular weight complexes. Following the interaction on the spectra of Pab1 RRM12 supports the second scenario. The 2D-TROSY spectrum reveals that the signals of the RRM1 domain remain unperturbed (i.e. not involved in interaction), therefore their faster T2 are attributable to an increase of the molecular weight of the complex (oligomer) (Figure 6). Further studies are required to address which of the three boxes are directly contacted by Pab1.
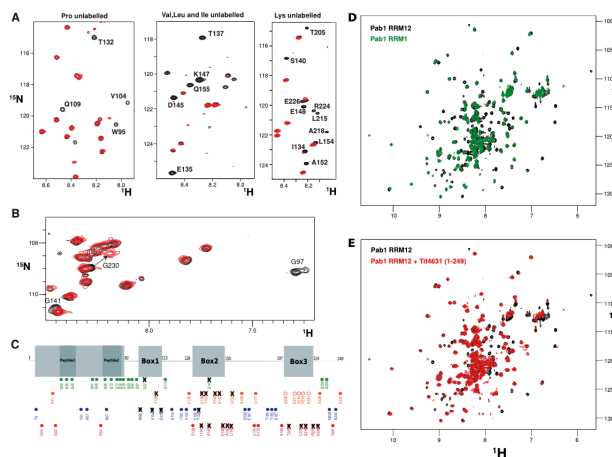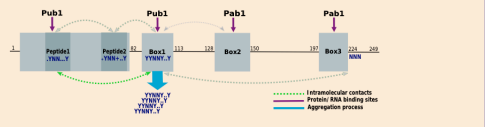

Figure 6. A) Isotopically discriminated ¹H-¹⁵N correlation spectra (IDIS). Tif4631 (1-249) reverse Pro and Lys samples were expressed using E. coli auxotrophic strains RF6 and RF10 in ¹³C/¹⁵N minimal media supplemented with unlabelled aminoacids². The standard E. coli BL21 (DE3) was used for reverse I,L,V by supplying unlabelled alfa-ketoacids. The IDIS spectra select signals following the unlabelled residue (i.e.¹²CO) and therefore map residues following Pro, Lys, Val, Ile and Leu. IDIS spectra of the complex with Pab1 RRM12 (unlabelled). **B)** Region of the Gly.**C)** Distribution of the mapped signals in the different spectra is shown on the scheme

Only Pab1 RRM2 is involved in Tif4631 recognition. **D)** Superposition of the NMR spectra of Pab1 RRM1 (green) and Pab1 RRM12 (black) (unstructured region preceding the RRM1 is included in both constructs). **E)** 2D TROSY spectra of the complex. Both spectra were acquired with ¹⁵N-labelled Pab1 RRM12 at concentration of 100 uM but the 2D TROSY spectra was recorded with 10-times more scans.

Conclusions

Tif4631 (1-249) contains up to 5 MoRFs which are sequence-conserved across different yeast strains. We showed that there are intramolecular interactions between some of them in the Tif4631 free state and that four of them are involved in interactions with the RRM domains of Pab1 (RRM2) and Pub1 (RRM3). Moreover, under certain circumstances, the system (Pub1-Pab1-Tif4631) can progress to an oligomer form. The BOX1 sequence is in some cases (perhaps in all) responsible for this behaviour and contains a sequence with high propensity to form amyloid-like structures. We propose a model in which BOX1 amyloid-like properties are critical for physico-chemical phase-transitions observed during the nucleation of stress granules (i.e. the formation of liquid droplets) and show how Pub1 and Pab1 interactions might regulate this process (Scheme).

References

1. Santiago, Martínez-Lumbreras. The role of Gbp2p, Nab2p and Pub1p along the mRNA cycle: structural and molecular recognition sudies by NMR and other biophysiscal techniques. (2013) PhD Thesis

2. Myat T. Lin, Risako Fukazawa, Yoshihary Miyajima-Nakano, Shinichi Matsushita, Sylvia K. Choi, Toshio Iwasaki, Robert B. Gennis. Escherichia coli Auxotroph Host Strains for Amino Acid Selective Isotope Labeling of Recombinant Proteins. (2015) Methods in Enzymology, Volume 565.

3. Joel R. Gillespie and David Shortle. Characterization of Long-range Structure in the DenaturedmState of Staphylococcal Nuclease I. Paramagnetic Relaxation Enhancement by Nitroxide Spin Labels. (1997) J. Mol. Biol. 268, 158-169.

4. Santiveri CM, Mirassou Y, Rico-Lastres P, Martínez-Lumbreras S, Pérez-Cañadillas JM. Pub1p C-terminal RRM domain interacts with Tif4631p through a conserved region neighbouring the Pab1p binding site. PLoS One. 2011; 6(9):e24481

5. Nozhat Safaee, Guennadie Kozlov, Anne M. Noronha, Jingwei Xie, Christopher J. Wills, and Kalle Gehring. Interdomain Allostery Promotes Assembly of the Poly(A) mRNA complex with PABP and eIF4G. (2012) Molecular Cell, 48, 375–386.

6. Rodolgo M. Rasia, Bernhard Brutscher and Michael J. Plevin. Selective Isotopic Unlabeling of Proteins Using Metabolic Precursors: Application to NMR Assignment of Intrinsically Disordered Proteins. (2012) ChemBioChem. 13, 732 – 739.

7. Golovanov AP, Blankley RT, Avis JM and Bermel W. Isotopically discriminated NMR spectroscopy: a tool for investigating complex protein

The N-terminal intrinsically disordered domain of eIF4G (Tif4631 in yeast) is a multi-tool for protein and nucleic acids recognition

Belén Chaves–Arquero[a], Santiago Martínez-Lumbreras[a,b] and Jose M. Pérez Cañadillas[a]

[a]Departamento de Química-Física Biológica. Instituto de Química-Física "Rocasolano". CSIC. C/Serrano 119, 28006 Madrid. Spain.
[b]Present address: Department of Chemistry. School of Biomedical Science. King's College London. 7 Trinity Street. London. SE1 1DB. United Kingdom.

"Scaffold" protein eIF4G1 (Tif4631 in yeast) forms a heterotrimeric complex with eIF4E (cap binding protein) and eIF4A (ATP-dependent RNA helicase) known as eIF4F that is crucial in mRNA translation initiation. The protein contains an N-terminals sequence for Pab1 recognition which facilitates the recircularization of the mRNA and promotes its early recruitment to the ribosomes. Several X-ray and NMR structures well characterize eIF4G-eIF4A and eIF4G-eIF4E complexes in both yeast and human systems.

However less is known for eIF4G1 large unstructured segments that are involved in other protein-protein and protein-RNA recognition activities.

Here we report the assignment of eIF4G1 1-249 (on the left) that contains several sites for crucial molecular interaction. To facilitate the assignment and the interaction analysis we made several selective unlabelled samples and used isotopically discriminated spectroscopy. For example in the IDIS spectra of the protein with Lys unlabeled (in red) only residues following Lys are detected. We also analysed the protein-protein or protein-RNA interactions using chemical shift mapping and paramagnetic relaxation enhancements (PREs).

Figure 1. Superposition of [1H,15N]-HSQC spectrum of eIF4G (1-249) (in black) in the Isotope-discrimianted spectrum of the selective Lys unlabelled sample (red). The distribution of the residues with preceding Lys is shown schematically in the model.

RNA - eIF4G interaction

Figure 4. Chemical shift mapping of different nucleotide length (RNA A8 and RNA A14) in eIF4G (1-82)

Figure 5. The [1H,15N]-HSQC spectrum of the eIF4G 1-82 construct (black) was compared with spectra of the same sample after titration with A14 RNA (red)
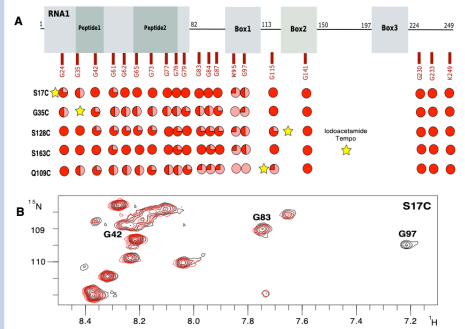
Mapping intramolecular interaction by PREs

Figure 2. A) Scheme of the PRE on the Gly signals for several eIF4G1 mutants. **B)** The superposition of [1H,15N]-HSQC spectra in reduced (red) and oxidized form (black) of TEMPO adduct.

Pub1 (RRM3) - eIF4G interaction

Figure 6. A) Overlay of selected regions of the [1H,15N]-HSQC spectra of eIF4G1 on its free (black) and Pub1 RRM3-bound forms (red). **B)** NMR chemical shift mapping (backbone amide groups) of the titration of Pub1 RRM3 with eIF4G1

Pab1 (RRM2) - eIF4G interaction

Figure 7. A) Isotopically discriminated 1H-15N correlation spectra (IDIS) of the complex Tif4631 1-249 and Pab RRM12 (unlabelled). The IDIS spectra select signals following the unlabelled residue (i.e. 13CO) and therefore map residues following Pro, Lys, Val, Ile and Leu. **B)** Regyon of the Gly. **C)** Scheme of the signals in the different spectra.

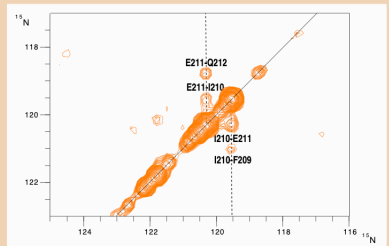## Discrete Secondary Structures

Figure 3. A plane of the double 15N-edited 3D-NOESY spectrum of eIF4G1 (1-249). The NH-HN spectrum exemplyfied alpha-helix in Box 3 regyon.

References

Selective aminoacids: Myat T. Lin, Risako Fukazawa, Yoshihary Miyajima-Nakano, Shinichi Matsushita, Sylvia K. Choi, Toshio Iwasaki, Robert B.Gennis.. Escherichia coli Auxotroph Host Strains for Amino Acid-Selective Isotope Labeling of Recombinant Proteins. (2015) Methods in Enzymology, Volume 565.
PREs: Joel R. Gillespie and David Shortle. Characterization of Long-range Structure in the DenaturedmState of Staphylococcal Nuclease I. Paramagnetic Relaxation Enhancement by Nitroxide Spin Labels. (1997) J. Mol. Biol. 268, 158-169.
Pub1: Santiveri CM, Mirasous Y, Rico-Lastres P, Martinez-Lumbreras S, Pérez-Cañadillas JM. Pub1p C-terminal RRM domain interacts with Tif4631p through a conserved region neighbouring the Pab1p binding site. PLoS One. 2011; 6(9):e24481
Pab1: Nozhat Safaee, Guennadie Kozlov, Anne M. Noronha, Jingwei Xie, Christopher J. Wilis, and Kalle Gehring. Interdomain Allostery Promotes Assembly of the Poly(A) mRNA complex with PABP and eIF4G. (2012) Molecular Cell. 48, 375–386.
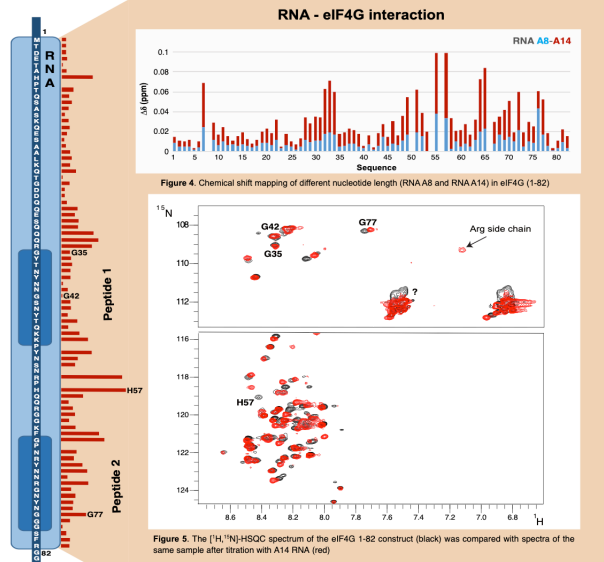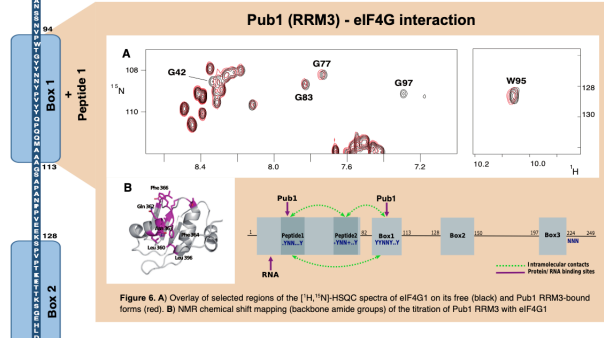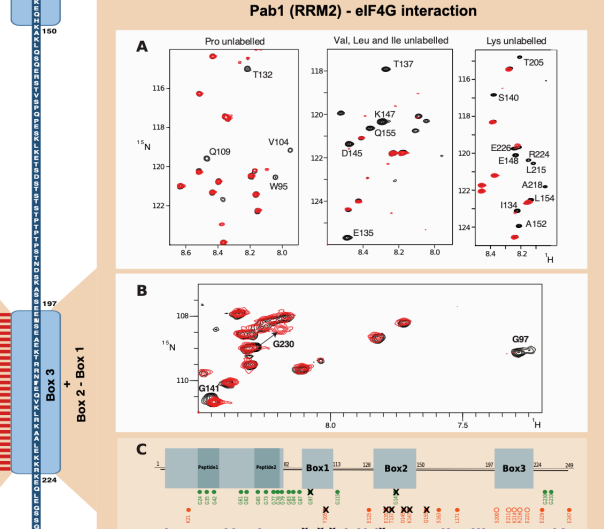
NMR studies of the intrinsically disordered C-terminal domain of Histone H1o: Effect of phosphorylation on structural, dynamic and DNA-binding properties

B. Chaves-Arquero[a], D. Pantoja-Uceda[a], A. Roque[b], I. Ponte[b], P. Suau[b], J.M. Pérez-Cañadillas[a] and M.A. Jiménez[a]

[a]Departamento de Química-Física Biológica, Instituto de Química-Física Rocasolano (IQFR). CSIC. C/Serrano 119, 28006 Madrid. Spain
[b]Departamento de Bioquímica y Biología Molecular, Facultad de Biociencias, Universidad Autónoma de Barcelona (UAB). Bellaterra, Barcelona. Spain

INTRODUCTION

Eukaryotes have a large percentage of intrinsically disordered proteins (IDP) which play key roles in biological processes.

The C-terminal domain of Histone 1 (C-H1.0) is an IDP, which is involved in the DNA binding and could be phosphorylated in (S/T)-P-X-(K/R)[1].

To analyse the effect of phosphorylation on the structure of C-H1.0, we characterised a construct with 3 sites of phosphorylation (pT22, pT44 and pT56) in its non-phosphorylated (C-H1.0) and tri-phosphorylated (pC-H1.0) states using a CON-based NMR assignment strategy.

To examine the effect of phosphorylation in DNA binding, we used NMR to determine the structures of a model peptide that contains a single phosphorylation site, both non- and phosphorylated in free state, and to examine their interaction with DNA.

RESULTS & DISCUSSION

1. C-H1.0 and pC-H1.0 assignments were achieved by a novel ^{13}C-detected CON-based strategy

C-H1.0 sequence contains 12 types of residues, three out of them represent about a 65% of the total (Figure 1A). As a consequence, the NMR spectra shows very little dispersion in the [1]H[N] nuclei (Figure 1B). We explored alternative assignment strategies based on ^{13}C-detected experiments[2]. Our proposal is based on the acquisiton of only 3D experiments: hacacoNcaNCO and hacaCOncaNCO and 2D [15]N-[13]CO (Figure 1C).
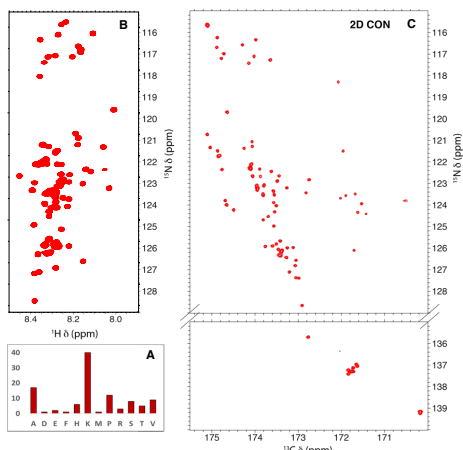

Figure 1. (A) Bar plot showing the total number of each residue type in the C-H1.0 sequence. The absent residues are not included in the plot. (B) 2D ^{1}H-^{15}N HSQC and (C) 2D [13]CO[15]N spectra recorded for C-H1.0 at 1 mM concentration in H2O/D2O 9:1 v/v at pH 5.5 and 25º C.

2. Structural and dynamics characterization of C-H1.0 and pC-H1.0

In both C-H1.0 and pC-H1.0, the chemical shift index (CSI)[3] and the chemical shift deviations of [13]Cα carbons ($\Delta\delta^{13}C\alpha = \delta^{13}C\alpha_{observed} - \delta^{13}C\alpha_{reference}$, ppm) do not show any tendency to form structure (Figure 2A).

Chemical shift comparation between C-H1.0 and pC-H1.0 (CSP) shows differences only in the three phosphorylated Thr (Figure 2B).

To obtain dynamic information we collected heteronuclear {[1]H}-^{15}N NOE experiment for both constructs and no significant differences can be appreciated (Figure 2C).
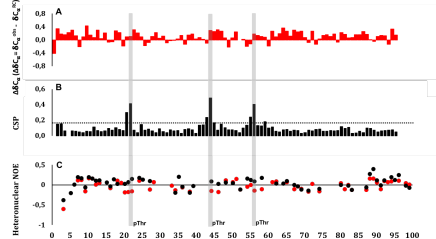

Figure 2. (A) Δδ[13]Cα values of C-H1.0 plotted as a function of residue number. Reference random coil values were taken from Poulsen[4]. (B) Chemical shifts perturbation (CSP) due to phosphorylation. (C) {[1]H}-^{15}N heteronuclear NOEs for C-H1.0 (red) and pC-H1.0 (black) in H2O/D2O 9:1 v/v at pH 5.5 and 25º C. The positions of the phosphorylated Thr are indicated by vertical grey bars.

REFERENCES

1. Talasz H1 et al. In vivo phosphorylation of histone H1 variants during the cell cycle. (1996). Biochemistry. Feb 13;35(6):1761-7.
2. Mäntylahti et al. Extension of the HA-detection based approach: (HCA)CON(CA)H and (HCA)NCO(CA)H experiments for the main-chain assignment of intrinsically disordered proteins. (2011). J Biomol NMR. 9;49:99–109
3. Noor E et al. CSI 3.0: a web server for identifying secondary and super-secondary structure in proteins using NMR chemical shifts, Nucleic Acids Research (2015): Vol. 43, Issue W1, p. W370-W377
4. Kjaergaard, M. et al. Random coil chemical shifts for intrinsically disordered proteins: Effects of temperature and pH (2011) J. Biomol. NMR 49(2):139-49.

Model peptide T44C-H1.0 or pT44C-H1.0

MDEPKRSVAFKKTKKEVKKVATPKKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKKPKVVKVKPVKASKPKKAKTVKPKAKSSAKRASVKKRSH

3. NMR structural study of the free peptides (T44C-H1.0 and pT44C-H1.0): Aqueous solution vs 90% TFE

The two peptides in aqueous solution and in 90 % TFE (Figure 3AB) show almost identical profiles. In aqueous solution, these peptides are mostly random coil, in equilibrium with a low population of helical structures (12-13 % at pH 5.5 and 5 ºC). Helical percentages increases greatly in 90 % TFE (53-55% at pH 5.5 and 5 ºC).

The differences in helix population between T44C-H1.0 and pT44C-H1.0 are in the range 1-4 % (experimental error 3-7%).
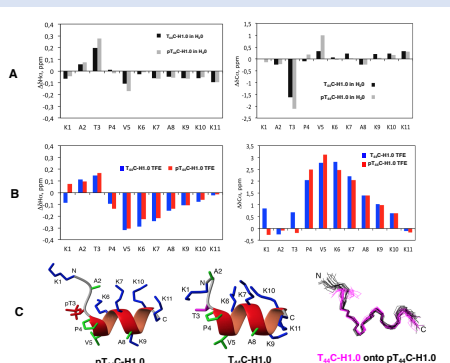

Figure 3. (A) Plots of [1]Hα and [13]Cα conformational shifts in aqueous solution for peptides T44-C-H1.0 (black) and pT44-C-H1.0 (grey) and (B) 90% TFE. (C) Representative structures of peptides T44-C-H1.0 and pT44-C-H1.0 in 90% TFE, and superposition of the backbone atoms of the two peptide ensembles.

4. DNA binding: NMR titrations

To get structural data about the DNA/peptide interaction, the titration of the two peptides with a 26 bp dsDNA was followed by NMR. Since the length and repetitive sequence of dsDNA make it too large its NMR spectra to be assigned, we analyzed the changes in the peptide NMR signals.

We mapped in both peptides the saturation point at [1 : 0.5] ratio. The chemical shift analysis between T44C-H1.0 and pT44C-H1.0 could indicate: (1) The residues next to pT44 (A43, T44 and V46) have a higher chemical shift. (2) In contrast, the residues affected for DNA interaction in T44C-H1.0 are K46, K50 and K51.
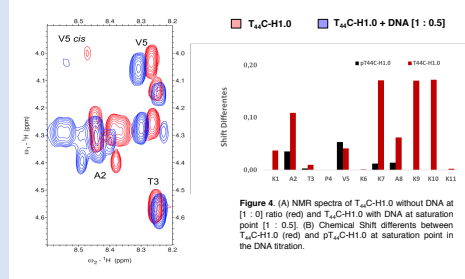

Figure 4. (A) NMR spectra of T44C-H1.0 without DNA at [1 : 0] ratio (red) and T44C-H1.0 with DNA at saturation point [1 : 0.5]. (B) Chemical Shift differents between T44C-H1.0 (red) and pT44C-H1.0 at saturation point in the DNA titration.

CONCLUSIONS

1. The higher dispersion of [13]C-^{15}N peaks relative to the [1]HN-^{15}N cross-peaks make the assignment of C-H1.0 and pC-H1.0 possible. Both proteins are very flexible and do not show any tendency to form structure in H2O/D2O 9:1 v/v at pH 5.5 and 25ºC.

2. We used two model peptides to study the effect of phosphorylation in the interaction with DNA. The two peptides T44C-H1.0 and pT44C-H1.0 are random coil in aqueous solution, but have a 55% of helical structure in 90 % TFE.

3. The interaction site of DNA in T44C-H1.0 seems to involve the lysines residues. However, the residues affected for DNA interaction in pT44C-H1.0 are A43, T44 and V46.

Multivalent interactions between eIF4G, Pab1 and Pub1 lead to the formation of phase-separated protein condensates

B. Chaves-Arquero[a], S. Zorrilla[b] and J.M. Pérez-Cañadillas [a]

[a]Departamento de Química-Física Biológica, Instituto de Química-Física Rocasolano (IQFR). CSIC. Madrid. Spain
[b]Departamento de biologia molecular y celular, Centro de Investigaciones Biologicas. CSIC. Madrid. Spain

Introduction

eIF4G1 forms a heterotrimeric complex with eIF4E (cap binding protein) and eIF4A (ATP-dependent RNA helicase) known as eIF4F that is crucial in mRNA translation initiation. Several and NMR structures well characterize eIF4G-eIF4A and eIF4G-eIF4E complexes.
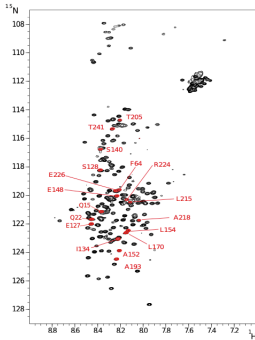
However less is known for eIF4G1 unstructured segment (1-249). We reported the assignment (on the left) that contains several sites for crucial molecular interaction. To facilitate the assignment and the interaction analysis we made several selective unlabeled samples and used isotopically discriminated spectroscopy. For example in the IDIS spectra of the protein with Lys unlabeled (in red) only residues following Lys are detected.
We analysed the protein-protein interactions (self recognition and RNA binding proteins interaction) using chemical shift mapping and paramagnetic relaxation enhancements (PREs). We also studied how protein-protein interaction modulate the formation of condensates using confocal microscopy.

Figure 1. Superposition of [1H,15N]-HSQC spectrum of eIF4G (1-249) (in black) in the isotope-discriminated spectrum of the selective Lys unlabeled sample (red). The distribution of the residues with preceding Lys is shown schematically in the model.

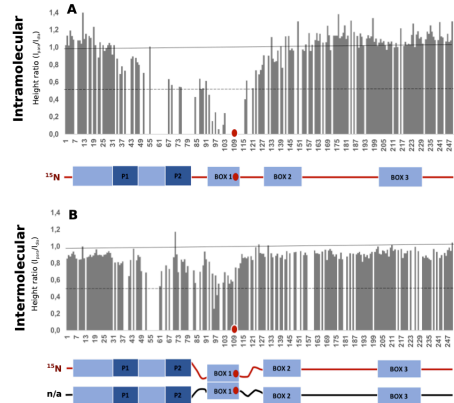## Mapping intramolecular and intermolecular interactions in eIF4G1 (1-249) by PREs

Figure 2. A) Paramagnetic relaxation enhancement of amide prontons in the spin-labeled eIF4G1 1-249 Q109C mutant (25 mM KPi, 25 mM NaCl, pH 6.5 and 25ºC), which was calculated form the peak intensity ratio of the HSQC spectra in the presence (paramagnetic) and absence (diamagnetic) of the spin label (red botton). B) eIF4G 1-249 + eIF4G Q109C in the presence and absence of the spin label at the same conditions.

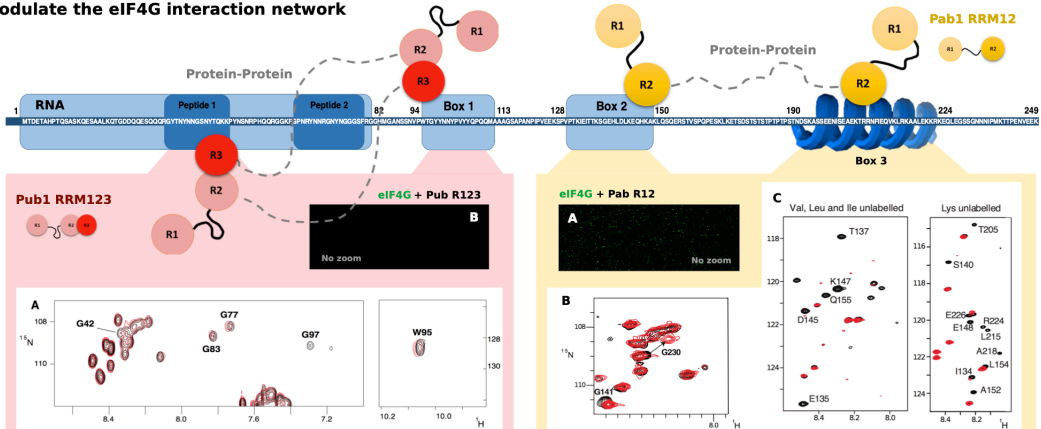## RNA binding proteins Pub1 and Pab1 modulate the eIF4G interaction network

Figure 3. A) Overlay of selected regions of the [1H, 15N]-HSQC spectra of eIF4G1 on its free (black) and Pub1 RRM3-bound forms (red). B) Confocal microscopy of the interaction between eIF4G 1-249 at 1uM (label in green) and Pub RRM123 at 5uM using Ficoll as crowder in PBS buffer.

Figure 4. A) Confocal microscopy of the complex eIF4G at 1uM (label in green) and Pab RRM12at 20uM using Ficoll as crowder in PBS buffer. B) Overlay of Gly regyon of the [1H, 15N]-HSQC spectra of eIF4G1 on its free (black) and Pab1 RRM12-bound forms (red). C) IDIS spectra of the complex eIF4G1 and Pab RRM12 (unlabelled). The IDIS spectra select signals following the unlabelled residue (i.e. 13CO) and therefore map residues following Lys, Val, Leu and Ile.

Conclusions

eIF4G contains up to 5 MoRFs which are sequence-conserved across different yeast strains. We mapped intra and intermolecular interactions in eIF4G free state and are involved in interactions with Pub1 and Pab1.
Moreover, under certain circumstances, the system (Pub1-Pab1-eIF4G) can progress to an oligomer form that can derivate to an condensates. At least Box1 sequence is involved in this behaviour.
We proposed a model in which the RBPs modulate the eIF4G intramolecular interaction network and this point is critical for physico-chemical phase-transitions observed during the nucleation of stress granules (i.e. the formation of condensates).

Insights into protein-protein condensates at condition simulation in vivo
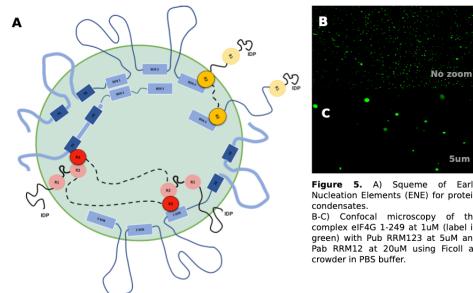
Figure 5. A) Squeme of Early Nucleation Elements (ENE) for protein condensates.
B-C) Confocal microscopy of the complex eIF4G 1-249 at 1uM (label in green) with Pub RRM123 at 5uM and Pab RRM12 at 20uM using Ficoll as crowder in PBS buffer.

References

Selective aminoacids: Myat T. Lin, Risako Fukazawa, Yoshihary Miyajima-Nakano, Shinichi Matsushita, Sylvia K. Choi, Toshio Iwasaki, Robert B.Gennis.. Escherichia coli Auxotroph Host Strains for Amino Acid-Selective Isotope Labeling of Recombinant Proteins. (2015) Methods in Enzymology, Volume 565.
PREs: Joel R. Gillespie and David Shortle. Characterization of Long-range Structure in the DenaturedmState of Staphylococcal Nuclease I. Paramagnetic Relaxation Enhancement by Nitroxide Spin Labels. (1997) J . Mol. Biol. 268, 158-169.
Pub1: Santiverl CM, Mirassou Y, Rico-Lastres P, Martinez-Lumbreras S, Pérez-Cañadillas JM. Pub1p C-terminal RRM domain interacts with Tif4631p through a conserved region neighbouring the Pab1p binding site. PLoS One. 2011; 6(9):e24481
Pab1: Nozhat Safaee, Guennadie Kozlov, Anne M. Noronha, Jingwel Xie, Christopher J . Wills, and Kalle Gehring. Interdomain Allostery Promotes Assembly of the PolyA) mRNA complex with PABP and eIF4G. (2012) Molecular Cell. 48, 375–386.

NMR studies on intrinsically disordered domains of Histone H1.0 and eIF4G: Effect of phosphorylation and multivalent protein-nucleic acid recognition

B. Chaves-Arquero, J.M. Pérez-Cañadillas and M.A. Jiménez

Departamento de Química-Física Biológica, Instituto de Química-Física Rocasolano (IQFR). CSIC. C/Serrano 119, 28006 Madrid. Spain

HISTONE H1.0

The C-terminal domain of Histone 1 (C-H1.0) is an IDP, which is involved in DNA binding and could be phosphorylated in (S/T)-P-X-(K/R)[1].

To analyse the effect of phosphorylation on the structural behaviour of C-H1.0, we characterised a construct with 3 sites of phosphorylation (pT22, pT44 and pT96) in its non-phosphorylated (C-H1.0) and tri-phosphorylated (pC-H1.0) states using a CON-based NMR assignment strategy[2].

Figure 1. Schematic representation of one conformer of C-terminal intrinsically disordered domain of Histone 1. The ensemble, of which one conformer is represented in red, were calculated with Flexible-Meccano algorithm using the sequence information.

RESULTS & DISCUSSION

1. C-H1.0 and pC-H1.0 assignments were achieved by a novel ¹³C-detected CON-based strategy
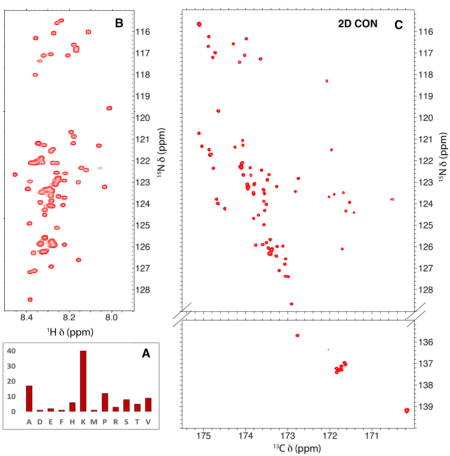

Figure 2. (A) Bar plot showing the total number of each residue type in the C-H1.0 sequence. The absent residues are not included in the plot. (B) 2D ¹H-¹⁵N HSQC and (C) 2D ¹³CO¹⁵N spectra recorded for C-H1.0 at 1 mM concentration in H₂O/D₂O 9:1 v/v at pH 5.5 and 25º C.

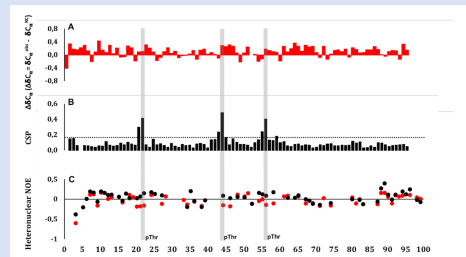### 2. Structural and dynamics characterization fo C-H1.0 and pC-H1.0

Figure 3. (A) Δδ¹³Cα values of C-H1.0 plotted as a function of residue number. Reference random coil values were taken from Poulsen[3]. (B) Chemical shifs perturbation (CSP) due to phosphorylation. (C) {¹H}-¹⁵N heteronuclear NOEs for C-H1.0 (red) and pC-H1.0 (black) in H₂O/D₂O 9:1 v/v at pH 5.5 and 25º C. The positions of the phosphorylated Thr are indicated by vertical grey bars.

CONCLUSIONS

1. The higher dispersion of ¹³C-¹⁵N peaks relative to the ¹Hᴺ-¹⁵N cross-peaks make the assignment of C-H1.0 and pC-H1.0 possible.

2. Both proteins are very flexible and do not show any tendency to form structure in H₂O/D₂O 9:1 v/v at pH 5.5 and 25ª C.

eIF4G

Eukaryotic translation initiation factor eIF4G plays a positive role in translation regulation by enhancing ribosome recruitment to the mRNA, but also a negative role by being an essential component of stress granules. The N-terminal disordered domain of eIF4G contains MoRFs involved in recognition of RNA binding proteins (Pub1 and Pab1), and in self-recognition.

Figure 1. Schematic representation of one conformer of N-terminal intrinsically disordered domain of eIF4G. This conformer (in blue) and the complete ensemble were calculated with Flexible-Meccano algorithm using the sequence information, chemical shifts and Residual Dipolar Couplings (RDCs).

RESULTS & DISCUSSION

1. Mapping intramolecular intractions in eIF4G (1-249) by Paramagnetic Relaxation Enhancement[1] (PREs)
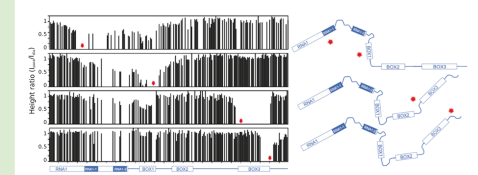

Figure 2. Paramagnetic relaxation enhancement of amide protons in the spin-labeled eIF4G. PRE data was calculated from the peak intensity ratio of the HSQC spectra in the presence (paramagnetic) and absence (diamagnetic) of the spin label (red start).

2. Loading poly(U) and poly (A) RNA binding proteins modulate the eIF4G interaction network
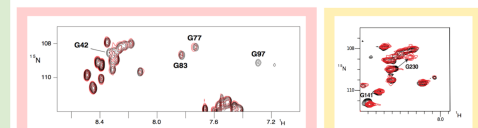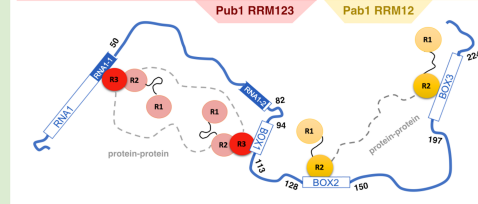

Figure 3. Overlay of Glycine region (G42 of RNA1-1 and G97 of BOX1) of the 2D ¹H-¹⁵N HSQC spectra of eIF4G on its free (black) and Pub1 RRM3-bound forms (red).

Figure 4. Region of the HSQC spectra of eIF4G on its free (black) and Pab1 RRM12-bound forms (red).

Pub1 RRM123

Pab1 RRM12

3. Insights into protein-protein condensates at condition simulation in vivo

Scheme of Nucleation Elements for protein condensates.

CONCLUSIONS

1. We mapped intra and intermolecular interactions in eIF4G free state and are involved in interactions with Pub1[2] and Pab1[3].

2. The system (Pub1-Pab1-eIF4G) can progress to an oligomer form that can derivate to condensates.

3. We proposed a model in which the RNA binding proteins modulate the eIF4G intramolecular interaction network and this point is critical for physico-chemical phase transitions observed during the nucleation of stress granules.
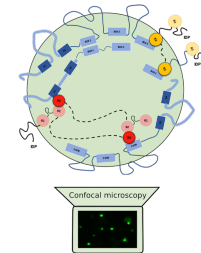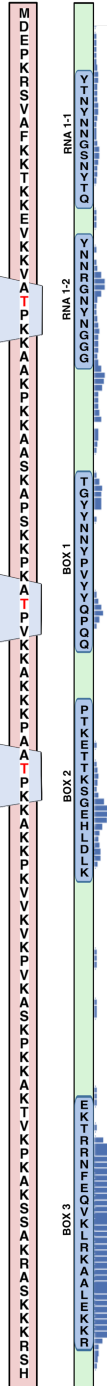

Confocal microscopy

Figure 5. Image of the complex eIF4G (label), Pub1 RRM123 and Pab1 RRM12 using Ficoll and PBS buffer.

REFERENCES

1. Talasz H1 et al. In vivo phosphorylation of Histone H1 variants during the cell cycle. (1996). Biochemistry. Feb 13;35(6):1761-7.
2. Chaves-Arquero B et al. A CON-based NMR assignment strategy for pro-rich intrinsically disordered proteins with low signal dispersion: the C-terminal domain of histone H1.0 as a case study. (2018). J. Biomolecular. NMR. 72:139-148.
3. Noor E et al. CSI 3.0: a web server for identifying secondary and super-secondary structure in proteins using NMR chemical shifts. (2015). Nucleic Acids Research. Vol. 43, Issue W1, p. W370-W377.

REFERENCES

1. Joel R. Gillespie and David Shortle. Characterization of Long-range Structre in the Denatured State of Staphylococcal Nuclease I. Paramagnetic Relaxation Enhancement by Nitroxide Spin labels. (1997) J. Mol. Biol. 268, 158-169.
2. Santiveri CM, Mirassou Y, Rico-Lastres P, Martínez-Lumbreras S, Pérez-Cañadillas JM. Pub1p C-terminal RRM domain interacts with Tif4631p through a conserved region neighbouring the Pab1p binding site. (2011). PLoS One. 6(9):e24481.
3. Nozhat Safaee, Guennadie Kozlov, Anne M. Noronha, J Ingwel Xie, Christopher J. Wills, and Kalle Gehring. Interdomain Allostery Promotes Assembly of the Poly(A) mRNA complex with PABP and eIF4G. (2012) Molecular Cell. 48, 375-386.

Sequence (central column):

RNA 1-1: M D E P K R S V A F K K T K K E V K K V A T P K K A A K P K K A A S K A P S K K P K A T P V K K A K K P A A T P K K A K K P K V V V K P V K A S K P K K A K T V K P K A K S S A K R A S K K K R S H

RNA 1-1: YTNYNGSNYTQ
RNA 1-2: YNNRGNYNGGG
BOX 1: TGYYNNYPVVYQPQQ
BOX 2: PTKETTSGEHLDLK
BOX 3: EKTRRNFEGVKLRKAALEKKR

V. ABBREVIATIONS

Abs	Absorbance	**INEPT**	Insensitive Nuclei Enhanced by Polarization Transfer
AcO	acetate	**IPTG**	Isopropyl β-D-1-thiogalactopyranoside
CD	Circular Dichroism	**IR**	Infrared
CDK	Cyclin-dependent kinase	**J**	Coupling constant
CSI	Chemical Shift Index	**LB**	Lysogenic Broth
CSM	Chemical Shift Mapping	**MALS**	Multi-angle light scattering
DIPSI	Decoupling In the Presence of Scalar Interactions	**MOPS**	3-(N-morpholino)propanesulfonic acid
DLS	Dynamic Light Scattering	**MoRF**	Molecular Recognition Feature
DMSO	Dimethylsulfoxide	**mRNA**	Messenger Ribonucleic Acid
DNA	Deoxyribonucleic Acid	**mRNP**	Messenger Ribonucleoprotein Particle
DSS	sodium 2,2-dimethyl-2-silapentane-5-sulphonate	**MW**	Molecular Weight
DTT	Dithiothreitol	**OD**	Optical Dispersion
EDTA	Ethylendiamintetraacetic acid	**NMR**	Nuclear Magnetic Resonance
eIF	eukaryotic Initiation Factor	**NOESY**	Nuclear Overhauser Effect Spectroscopy
FPLC	Fast protein liquid chromatography	**PAGE**	Polyacrilamide Gel Electrophoresis
H1	Histone H1 protein	**PABPs**	Poly (A) Binding proteins
HSQC	Heteronuclear Simple Quantum Correlation	**PBS**	Phosphate Buffered Saline
IDD	Intrinsically Disordered Domain	**PEG**	Polyethilenglycol
		PIPES	Piperazine-N,N'-bis(2-ethanesulfonic acid)
IDP	Intrinsically Disordered Protein	**PCR**	Polymerase Chain Reaction

PUBP Poly (U) binding protein

PRE Paramagnetic Relaxation Enhancement

RDC Residual Dipolar Coupling

RNA Ribonucleic Acid

RRM RNA Recognition Motif

SAXS Small angle X-ray scattering

SG Stress Granules

SDS Sodium Dodecyl Sulphate

TEV Tobacco Etch Virus

TOCSY Total Correlation Spectroscopy

TxA Thiorredoxine A

UV Ultraviolet