Universitat Autònoma de Barcelona

**DEPARTAMENT DE BIOQUÍMICA I BIOLOGIA MOLECULAR**

INSTITUT DE RECERCA BIOMÈDICA
IRB BARCELONA

**PROGRAMA DE BIOLOGIA ESTRUCTURAL I COMPUTACIONAL**

# STRUCTURAL STUDIES ON FF DOMAINS BY NMR SPECTROSCOPY

Roman Bonet Figueredo

2009

Universitat Autònoma de Barcelona

**DEPARTAMENT DE BIOQUÍMICA I BIOLOGIA MOLECULAR**

INSTITUT DE RECERCA BIOMÈDICA
IRB BARCELONA

**PROGRAMA DE BIOLOGIA ESTRUCTURAL I COMPUTACIONAL**

# STRUCTURAL STUDIES ON FF DOMAINS BY NMR SPECTROSCOPY

Thesis presented by Roman Bonet Figueredo to apply for the degree of PhD in Biotechnology from the Universitat Autònoma de Barcelona

This work has been done in the Protein NMR group from the Institute for Research in Biomedicine of Barcelona, under the supervision of Dr. Maria J. Macías.

Roman Bonet Figueredo

With the agreement of the Thesis Supervisor

Dr María J. Macías Hernandez

Barcelona, Juliol de 2009

## AGRAÏMENTS

Des d'aquí m'agradaria expressar el meu agraïment a tota la gent que, d'alguna manera o altra, m'ha ajudat en el llarg recorregut que suposa la realització de la tesi doctoral.

Primerament, vull agrair a la meva directora de tesi, la Dra. María J. Macías, haver-me permès realitzar aquest treball en el seu grup. Pero sobretot vull donar-li les gràcies pel seu suport al llarg de tot aquest temps així com per les facilitats que m'ha donat per a poder desenvolupar el projecte presentat en aquesta tesi doctoral.

Però, el meu camí en el món del laboratori i del doctorat havia començat abans de ser en el grup de la María, i per això m'agradaria també recordar a tota la gent que m'ha acompanyat durant aquests anys: per començar, vull agrair al Dr. Xavier Avilés i al Dr. Josep Vendrell que em donessin l'oportunitat d'entrar en el seu grup d'investigació. Agraeixo als meus companys PROs i també als d'Enzimo de l'IBB tot el que em van ensenyar en aquell període. I no em vull oblidar de la resta de companys del Departament, pel bon ambient i els moments compartits durant tot aquell temps.

Ja aquí al Parc Científic, vull fer referència a tots els meus companys de grup amb qui he compartit laboratori durant els últims quasi 5 anys. Alguns d'ells ja han marxat, com la Ximena, la Begonya o el Zhang Lei, i altres encara són per aquí, com la Lídia, el Pau, l'Eric, la Nina o el Tiago. Vull fer una menció especial a la Ximena, per tota la seva ajuda quan vaig arribar al laboratori, i al Pau per haver-me introduït en l'extrany món de la RMN. Però, sobretot m'agradaria agrair a la Lídia el seu suport en una part important de l'experimental realitzada al llarg d'aquesta tesi. I per suposat, aquí també he conegut força gent amb qui he compartit bones estones, com són la gent del grup de RMN de biomolecules i tota la secció de cristalògrafs.

Finalment, vull agrair el suport tant important dels meus pares, la meva família i els meus amics amb els quals sempre he pogut comptar quan els he necessitat. I vull donar les gràcies específicament a dues persones molt importants en la meva vida, el Dani i la Marina. Vull que sapigueu que sense vosaltres possiblement aquest treball no l'hauria pogut tirar endavant i que sempre que ho necessiteu em tindreu al vostre costat.

# RESUM

El present treball s'ha realitzat en el marc del Programa de Biologia Estructural i Computacional de l'IRB Barcelona, en el grup de RMN de proteïnes, sota la direcció de la Dra María J. Macías. L'interès del grup d'investigació se centra en la determinació estructural de dominis de proteïna, en l'estudi d'interaccions i en entendre els mecanismes per a la seva regulació. Concretament, aquest treball està enfocat en l'estudi estructural de dominis FF.

En el nostre grup es va començar a estudiar els dominis FF a partir de l'observació que aquests dominis es troben sovint en proteïnes que també contenen WW, un petit domini d'interacció proteïna-proteïna que ha estat i continua sent la línia d'investigació principal del grup, i del qual s'han realitzat nombrosos i detallats estudis.

En canvi, el dominis FF han estat identificats fa poc temps i la informació estructural i funcional de la qual es disposa és, a dia d'avui, escassa. En part, el pocs estudis realitzats amb aquests dominis pot ser conseqüència de que només els trobem en un conjunt reduït de proteïnes. De fet, la seva presència es limita a tres famílies de proteïnes: els factors d'splicing FBP11, Prp40 i URN1, el factor de transcripció CA150 i una família de proteïnes reguladores de RhoGTPases, les p190 RhoGAPs.

Al meu grup, l'estudi amb dominis FF es va iniciar amb la resolució de l'estructura tridimensional del primer FF de la proteïna Prp40 de llevat, que va representar la segona estructura resolta per a un domini FF. En aquest treball l'objectiu ha estat aprofundir en el coneixement dels dominis FF, bàsicament des d'un punt de vista estructural utilitzant com a tècnica principal la resonànica magnètica nuclear (RMN).

Aquesta tesi doctoral s'ha dividit en tres parts i en cada una d'elles s'ha treballat amb els dominis FF de les tres famílies de proteïnes esmentades anteriorment. L'enfoc també ha estat lleugerament diferent a cada part. Així, en el primer capítol, centrat en els dominis FF presents als factors d'splicing URN1 i Prp40, el gruix de la feina va ser l'obtenció de noves estructures tridimensionals per RMN d'aquests dominis. En canvi, en el segon capítol ens vam centrar en l'estudi de la interacció dels dominis FF de CA150 amb el seu primer lligand descrit, un motiu doble fosforilat del domini C-terminal de la RNA-polimerasa II (fosfo-CTD). Finalment, en la tercera part l'interès es va dirigir a l'estudi de la regulació de l'associació dels dominis FF de p190-A RhoGAP amb el factor de transcripció TFII-I, per mitjà d'una fosforilació sobre el primer FF de p190-A RhoGAP.

En la primera part del treball, es va poder determinar l'estructura tridimensional per RMN de l'únic domini FF de URN1 i del domini FF4 de Prp40. Ambdós mostren la clàssica arquitectura $\alpha1-\alpha2-3_{10}-\alpha3$ descrita anteriorment per aquest dominis, que demostra que tot i la baixa identitat seqüencial entre els diferents dominis FF, el plegament global es troba conservat.

També es va realitzar una comparació de les superfície electrostàtiques dels dominis FF per analitzar la implicació de la distribució de càrregues en el reconeixement de lligand. Fins aquest treball es considerava que el valor global de $pK_a$ (altrament referit com a punt isoelèctric, pI) era un bon indicador de l'especificitat de lligand per als dominis FF, però nosaltres hem comprovat que la distribució de càrregues no es troba conservada entre dominis tot i que aquests tinguin valors similars de $pK_a$. De fet, es va verificar aquesta observació amb la confirmació que el domini FF de URN1, amb un valor global de $pK_a$ molt similar al de FF1 de Prp40, no interacciona amb el motiu TPR descrit com a lligand per aquest últim.
També, en aquesta primera part es van realitzar comparacions estructurals amb altres proteïnes que presenten un plegament similar al dels dominis FF amb la idea de detectar una possible conservació de regions d'unió a lligand a nivell estructural i es va observar que les zones que inclouen les hèlixs $\alpha 2$ i $3_{10}$ són les regions més variables. Remarcablement, en aquestes proteïnes en general la segona hèlix es troba implicada en el reconeixement de lligand.

El treball realitzat en la segona part va consistir en el mapeig de les superfícies d'interacció dels dominis FF de CA150 per a la unió prèviament descrita amb el fosfo-CTD. En concòrdancia amb les observacions realitzades en el primer capítol, els residus implicats en aquestes interaccions es troben situats bàsicament en regions que comprenen el primer gir i les hèlixs $\alpha 2$ i $3_{10}$. Aquesta zona d'unió, però, no coincideix amb la regió mapejada al domini FF1 de FBP11 per a la interacció amb el mateix lligand i que es localitza a les hèlixs $\alpha 1$ i $\alpha 3$.
Els resultats d'aquesta part han confirmat, d'acord amb el que s'havia comprovat en altres treballs, que la interacció dels dominis FF de CA150 amb els seus lligands presenta una afinitat baixa, i que no sembla existir un efecte cooperatiu en les interaccions.

La feina realitzada a l'ultima part d'aquesta tesi ha permès conèixer un mecanisme particular de fosforilació per al domini FF1 de p190-A RhoGAP, així com característiques estructurals diferents per a aquest domini en comparació a altres dominis FF. El domini FF1 de p190-A RhoGAP no presenta la clàssica hèlix $3_{10}$ sinó que és un domini integrament format per hèlixs $\alpha$. A més, es va observar que contactes del primer gir amb residus addicionals a l'extrem C-terminal són indispensables per a l'estabilitat i el correcte plegament del domini. També es va observar que l'estabilitat del domini representa un factor clau per a la fosforilació, ja que es va comprovar que el desplegament previ del domini FF era necessari per a que es produís la fosforilació. Així, sembla que la capacitat de fosforilar-se d'aquest domini FF es troba estretament lligada amb canvis conformacionals en la seva estructura terciària.

# TABLE OF CONTENTS

# INDEX OF FIGURES AND TABLES

## FIGURES

## TABLES

# ABBREVIATIONS

**Å** ..... Ångström
**APBS** ..... Adaptative Poisson-Boltzmann Solvent
**ATP** ..... Adenosine Triphosphate
**Bo** ..... External magnetic field
**BLAST** ..... Basic Local Alignment Search Tool
**BMRB** ..... Biological Magnetic Resonance Bank
**$^{13}C$** ..... carbon-13 isotope
**CTD** ..... C-terminal Domain of the RNA polymerase II
**δ** ..... chemical shift
**Da** ..... Dalton
**DTT** ..... Dithiothreitol
**FF** ..... FF domain (contains two conserved Phe residues)
**FID** ..... Free Induction Decay
**Fmoc** ..... 9-fluorenylmethyloxicarbonyl
**FT** ..... Fourier Transform
**GST** ..... Glutathione-S-Transferase
**$^{1}H$** ..... hydrogen-1 nucleus
**HPLC** ..... High Performance Liquid Chromatography
**HSQC** ..... Heteronuclear Single Quantum Coherence
**IPTG** ..... Isopropyl-β-D-thiogalactopyranosid
**MALDI-TOF** ..... Matrix-Assisted Laser Desorption/Ionization-Time-Of-Flight
**MS/MS** ..... Tandem Mass Spectroscopy
**MRI** ..... Magnetic Resonance Imaging
**$^{15}N$** ..... nitrogen-15 isotope
**NMR** ..... Nuclear Magnetic Resonance
**NOE** ..... Nuclear Overhauser Effect
**NOESY** ..... Nuclear Overhauser Effect Spectroscopy
**PCR** ..... Polymerase Chain Reaction
**PDB** ..... Protein Data Bank
**PDGF** ..... Platelet Derived Growth Factor
**phospho-CTD** ..... Hyperphosphorylated CTD (on Ser 2 and 5 of the consensus sequence)
**pI** ..... Isoelectric point
**$pK_a$** ..... logarithmic measure of the acid dissociation constant
**ppm** ..... parts per million
**RDC** ..... Residual Dipolar Coupling
**RhoGAP** ..... Rho GTPase Activating Protein
**RMSD** ..... Root Mean Square Deviation
**SCOP** ..... Structural Classification Of Proteins database
**SDS-PAGE** ..... Sodium Dodecyl Sulphate - Poliacylamide Gel Electrophoresis
**SGD** ..... Saccharomyces Genome Database
**$T_1$** ..... longitudinal (spin-lattice) relaxation
**$T_2$** ..... transverse (spin-spin) relaxation
**TEV** ..... Tobacco Etch Virus

**TFA** ..... Trifluoroacetic Acid
**TIS** ..... Triispropylsilane
**TMS**..... tetramethylsilane
**TOCSY** ..... Total Correlation Spectroscopy
**TROSY** ..... Transverse Relaxation Optimization Spectroscopy
**TPR** ..... Tetratricopeptide Repeat
**TRX**..... Thioredoxin
**wt** ..... wild type

# GENERAL INTRODUCTION

# PART 1- NMR SPECTROSCOPY

The first $^{1}H$ NMR (Nuclear Magnetic Resonance) spectrum of a protein was obtained in 1957 for the ribonuclease in a 40 MHz spectrometer (Saunders, 1957). Since then, the importance of NMR in the field of biological science has increased constantly. Nowadays, NMR spectroscopy is used for the structure determination of biomolecules in solution or in solid-state as well as for the study of their physical properties, interactions and dynamics. In the field of medicine, the MRI (Magnetic Resonance Imaging) is a common, NMR-based technique for obtaining detailed images of the body.

NMR impact in many fields increased from two major technical advances: the introduction of the pulsed Fourier Transform (FT) NMR and multi-dimensional NMR spectroscopy.

Other technical improvements have been the progressive increase of the magnetic fields (up to 950 MHz for proton nowadays), the appearance of multichannel spectrometers, the introduction of pulsed gradients or methodologies such as Transverse Relaxation Optimization Spectroscopy (TROSY) experiments and residual dipolar couplings (RDCs).

## 1.1 The Magnet

A NMR spectrometer is, in fact, a sort of Dewar glass that needs to keep a very low temperature, around 4K, in order to maintain the properties of superconductivity of the solenoid, which allows the generation of the magnetic field. The solenoid is formed by alloys based on niobium, which are immersed in a bath of liquid helium. This inner Dewar is contained within an outer jacket of stainless still or aluminium that contains liquid nitrogen. $N_2$ and He should be refilled in the magnet periodically, every week in the case of $N_2$ and every 2-9 months in the case of the He (depending on the model).

The NMR probe is the element of the magnet that transmits the radiofrequency pulses to the sample and receives the generated signal. The NMR probe is placed in a hollow tube, at room temperature, in the centre of the magnet, and is introduced in it from the bottom to the active region of the magnet. In general, the probe can be exchanged depending on the experiment to perform. Around the probe is mounted a set of gradient coils, termed the shims.

The samples for the magnet are held in thin cylindrical tubes and are introduced in the magnet from the top of the hollow tube. Then, from the top of the magnet the sample is placed in the probe by means of an air current.

The NMR probe-head is connected to, at least, three radio-frequency cables to provide the $^{2}H$ lock (a signal used as a measurement of the homogeneity in the magnetic field), the $^{1}H$ frequency and the hetero-nucleus frequency.

The magnet also incorporates other devices such as a heater to control the temperature, or the air system to lift the sample inside and outside the magnet.

A picture and a scheme of a superconducting magnet are depicted in Fig. 1.

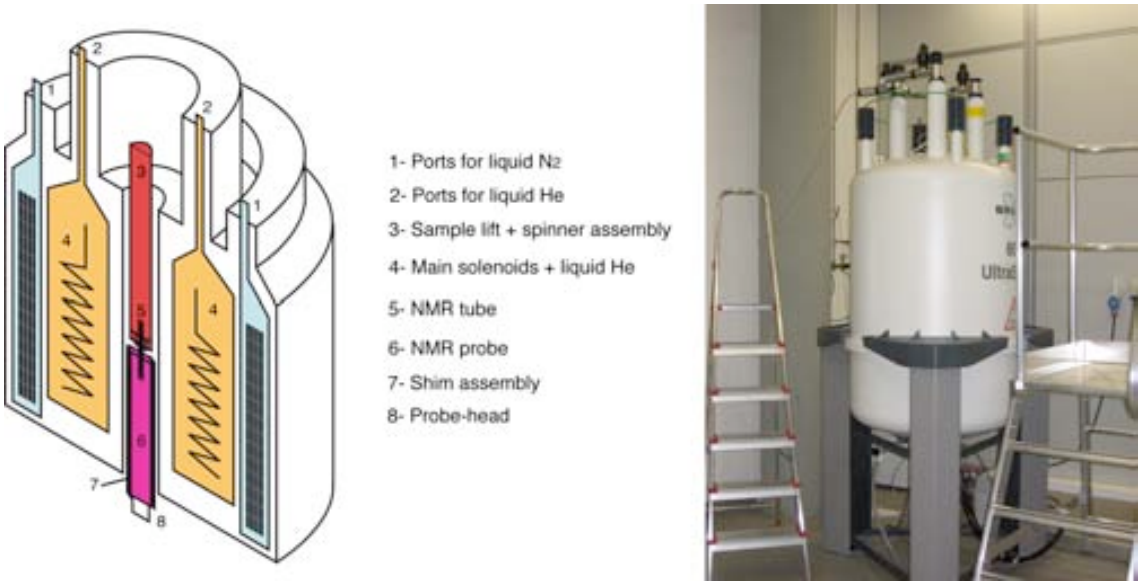



**Figure 1: The magnet**

*Left*- scheme of a superconducting magnet with the main parts displayed *Right*- Bruker Advance III-600 magnet at the IRB

## 1.2 Basic principles of NMR spectroscopy

NMR spectroscopy is based on an energy absorption by the nuclei of the molecule. However, the absorption can only take place if the nuclei possess a magnetic moment (μ). The magnetic moment is a quantum mechanical property of the nucleus that arises from another intrinsic nucleus property, the spin (I)

$$\vec{\mu} = \gamma \hbar \hat{I}$$

where $\gamma$ is the gyromagnetic ratio, another inherent property of the nucleus that can be regarded as its sensitivity to NMR.

The spin of a given nucleus depends on its charge / mass equilibrium and thus, some nucleus have values of I = 0 and others I ≠ 0. For example $^{1}H$ have to possible spin states, I = ½ and I = -½.

Notably, only nuclei with I ≠ 0 can be observed by NMR. When an external magnetic field ($B_o$) is applied, the different μ of a nucleus adopt distinct orientations, parallel or anti-parallel to the

$B_o$, each one with a different energy level. The magnitude of the $B_o$ applied also defines de distance between the energy levels, as depicted in Fig. 2:

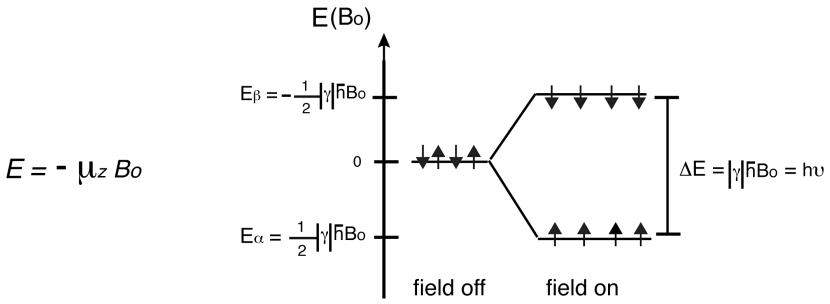



**Figure 2: Effect of a magnetic field on the nuclear spins.**

Nuclear spin energy levels in absence and presence of a static magnetic field. Transitions between states are possible when the energy of electromagnetic radiation matches de resonance frequency of the nuclei.

Moreover, spin nuclei under a magnetic field experience a rotational movement named precession. For a given nucleus, the angular frequency at which it precesses is known as the Larmor frequency, and is defined by:

$$\nu_o = \frac{\gamma B_o}{2\pi}$$

Resonance absorption by the nuclei only occurs if the incident electromagnetic radiation has the same frequency as the Larmor frequency of the nucleus (Fig. 2).
Remarkably, the difference between the number of spins in the two states is very small, even under strong external magnetic fields, and only this difference contribute to the NMR signal. Thus, NMR is an insensitive technique compared to other spectroscopic methods.

### 1.2.1 Radio frequency pulses

To obtain a NMR spectrum, the system has to be brought in a non-equilibrium state to allow the occurrence of transitions between the two spin states. In the equilibrium state, there is a net macroscopic magnetization ($M_o$) along the z-axis resulting from the precession of the nuclei at the resonance frequency. After the introduction of a short radio frequency pulse, that covers the range of precession frequencies of the nuclei, the equilibrium is perturbed and the $M_o$ deviates towards the xy plane, causing the appearance of transverse magnetization (Fig. 3). This transverse magnetization induces a radio-frequency current that is detected in the receiver coil and converted into a measurable signal.

However, spins lost their synchrony over time due to relaxation, and this results in a progressive disappearance of the transverse magnetization and, therefore, of the signal. This phenomenon occurs by means of two different mechanisms ($T_1$ and $T_2$).

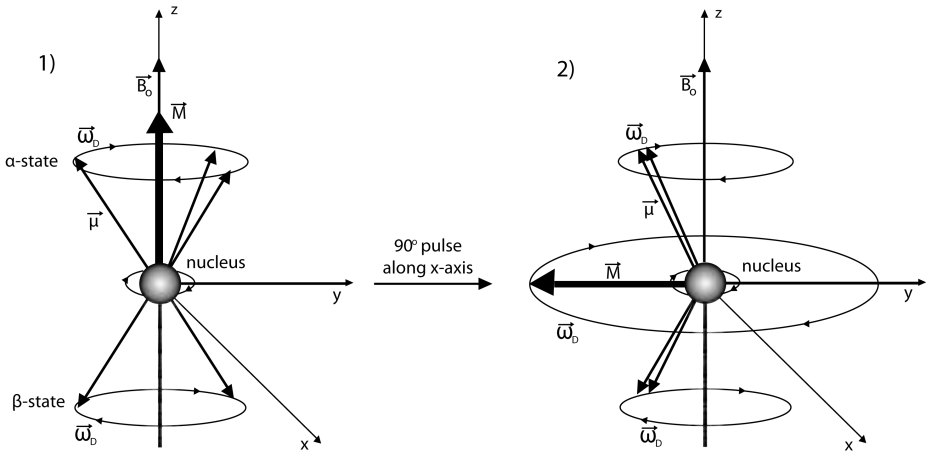



**Figure 3**: **Effect of a 90º pulse on the magnetisation.**

1) In the equilibrium state, the precession of the nuclei under the magnetic field ($B_o$) generates a net magnetisation $M_o$ (longitudinal magnetisation) along the z-axis.

2) After a 90º pulse along the x-axis, the equilibrium is perturbed and spins precess in a synchronized form, causing the appearance of transverse magnetisation in the xy plane.

The resonance signal that we measure is referred as free induction decay (FID), and it can be understood as a measure of the ability of spins to return to the equilibrium. Since all spins precess at slightly different resonance frequencies, the FID is a composition of multiple signals as a function of time. The subsequent processing by Fourier Transformation of the FID yields a spectrum in the frequency domain, the NMR spectrum.

In NMR experiments, the majority of radio frequency pulses used are 90º pulses or 180º pulses that turn the macroscopic magnetisation by 90º (as depicted in Fig. 3) or by 180º respectively.

### 1.2.2 Chemical shift

The term 'chemical shift' refers to the frequency at which a given nucleus resonates. One could think that, as all atoms in the molecule experience the same magnetic field, all spins of a given nucleus resonate at the same frequency. But this is not the case, as the resonance frequency of a nucleus depends on the effective magnetic field that it experiences, and this depends on the local chemical environment for that nucleus.

The resonance frequency value for a nucleus *i* is obtained from the next formula:

$$\nu_i = -\frac{\gamma B_0}{2\pi}(1-\sigma_i)$$

where $\nu_i$ is the resonance frequency of the nucleus *i*, $\gamma$ is the gyromagnetic ratio of the nucleus, $B_o$ is the actual strength of the magnet and $\sigma_i$ is the average shielding for the nucleus *i*.

The chemical shift is, in contrast, a field-independent magnitude. It is reported relative to a reference frequency $\nu_{ref}$. For $^1H$ and $^{13}C$, tetramethylsilane (TMS) is commonly used as the reference. So, the difference between the frequency of the signal and that of the reference, divided by the frequency of the reference gives the chemical shift for a signal. As the frequency shift values for the signals are very small compared to the applied frequency of magnetic field, chemical shifts are expressed in parts per million (ppm), as reflected in the formula:

$$\delta_i^{\text{ppm}} = \frac{\nu_i - \nu_{\text{ref}}}{\nu_{\text{ref}}} \cdot 10^6$$

The shielding $\sigma_I$ for a given nucleus depends on the electrons surrounding this nucleus. The circulation of the electrons generates and additional induced secondary field that opposes to the applied magnetic field and causes, for a nuclei *i*, to experience an effective field different from the applied field and different from the field experienced for another nucleus *j* with a different configuration of electrons around. Overall, this is translated in different chemical shifts for nucleus *i* and *j*.

The chemical shift of a nucleus *i* is influenced by factors such as the electronegativity of the neighbouring groups, or anisotropic effects.

For example, generally, protons close to oxygen groups have greater chemical shifts than protons attached to carbon groups, as oxygen is a more electronegative atom and reduces the electron density surrounding the proton, that is, therefore, more deshielded.

The anisotropic effects account for the induced magnetic fields generated by electrons in double bonds or aromatic systems. In this way, we find protons of aromatics rings with shifts around 6-7 ppms whereas protons of methyl groups appear around 1-2 ppms.

In summary, we can, roughly, identify the functional groups of a molecule on the basis of the chemical shifts observed for the signals. Moreover, chemical shifts are used to monitor changes in the molecule induced by temperature, pH or ligand binding.

### 1.2.3 J-coupling

The J-coupling or scalar coupling arises from the interaction between different nuclei in a molecule. Nuclei that are correlated through this kind of interactions are grouped into spin systems.

The J-coupling does not depend on the external $B_o$ but on the chemical characteristics of the molecule. The interaction occurs through the chemical bonds of the molecule, and could be understood as the transferring of magnetisation from one spin to another, resulting in the

splitting of NMR signals. The J-coupling can be homonuclear or heteronuclear, and is the responsible of the presence of signal multiplicity in NMR spectra. To observe J-coupling between two nuclei they should have different chemical shifts. In addition, the value of the three-bond J-coupling constant (or, in other words, the magnitude of splitting of the signal), $^{3}J$, depends on the torsion angles between the coupled spins, and thus yields valuable dihedral angle information for structure calculation.

In proteins, long-range couplings (more than 3 bonds apart in the case of $^{1}H$) are generally too small to cause observable splittings. Hence, only spin systems within individual amino acids can be obtained in proton spectra. If the protein is $^{15}N$, $^{13}C$-labelled, J-couplings between $^{1}H$, $^{15}N$ and $^{13}C$ allow correlations of nuclei across the peptide bond. These correlations between nuclei are the basis of correlation spectroscopy (COSY) NMR experiments.

### 1.2.4 NOE (Nuclear Overhauser Effect)

The Nuclear Overhauser Effect (NOE) is also a magnetisation transfer from one spin to another. However, in this case the transfer occurs through the space and arises from direct dipole-dipole interactions in a phenomenon named cross-relaxation. Particularly, the NOE is the change of intensity of a signal due to the dipolar interaction. In addition, this intensity change of the signal has a direct correlation with the distance between the two spins involved. Typically, a distance of 5-6 Å is the limit for detecting a NOE peak between two protons.

Consequently, NOEs yield distance information of the atoms of the molecule and are a most valuable data for the structure determination.

### 1.2.5 Relaxation

In NMR spectroscopy, the term relaxation applies to the return of the nuclear magnetisation to the equilibrium state. There are two different physical mechanisms by which relaxation occurs, termed $T_1$ (longitudinal or spin-lattice relaxation) and $T_2$ (transverse or spin-spin relaxation).

The reason for the study of relaxation processes is that they correlate with structural features of the molecules such as their internal motions. Another interesting aspect of relaxation study is that it has greatly helped in the design of NMR experiments, for example with the introduction of spin-echoes.

$T_1$ relaxation accounts for the rate at which the longitudinal magnetization (Mz) recovers. $T_1$ values strongly depend on NMR frequency and thus vary with the external $B_o$. $T_1$ mechanisms involve the dissipation of energy by interaction of a spin with the surrounding nuclei. Thus, $T_1$ also depends on the mobility of the molecule. $T_1$ has rather long values and always $T_1 > T_2$.

$T_2$ relaxation accounts for the rate of disappearance of transverse magnetisation, due to the loss of coherence of precessing spin nuclei. To this intrinsic relaxation process, however, we have to add the loss of coherences that result from the inhomogeneity of the $B_o$ and thus we refer to $T_2$* to describe the overall tranverse relaxation. $T_2$* can also be correlated with the mobility of the molecules and with the width of NMR signals: for molecules with high molecular weights ($M_w$), $T_2$* are shorter than for small molecules. Thus, relaxation process is faster and this is translated in wider bands for the NMR signals.

## 1.3 Protein structure determination by NMR

X-ray crystallography is widely used for protein structure determination. As observed in Fig. 4, in 2008, the number of solved-structures by X-ray crystallography has been 10-times greater than the number protein structures determined by NMR. In absolute terms, the number of X-ray structures is about 6 times the number of NMR structures.

In general, protein structure determination by NMR is limited by the size of the protein, since big proteins (over ~30 kDa) result in crowded and ambiguous spectra that difficult data analysis. However, the development of higher field magnets and new pulse sequences such as TROSY (reviewed in (Fernandez & Wider, 2003)), or the incorporation of new isotope labelling techniques (Staunton et al, 2006)) is helping to overcome these classical problems. Besides, NMR is a powerful tool for investigating protein-ligand interactions, molecular dynamics or structural rearrangements upon modifications such as phosphorylation.

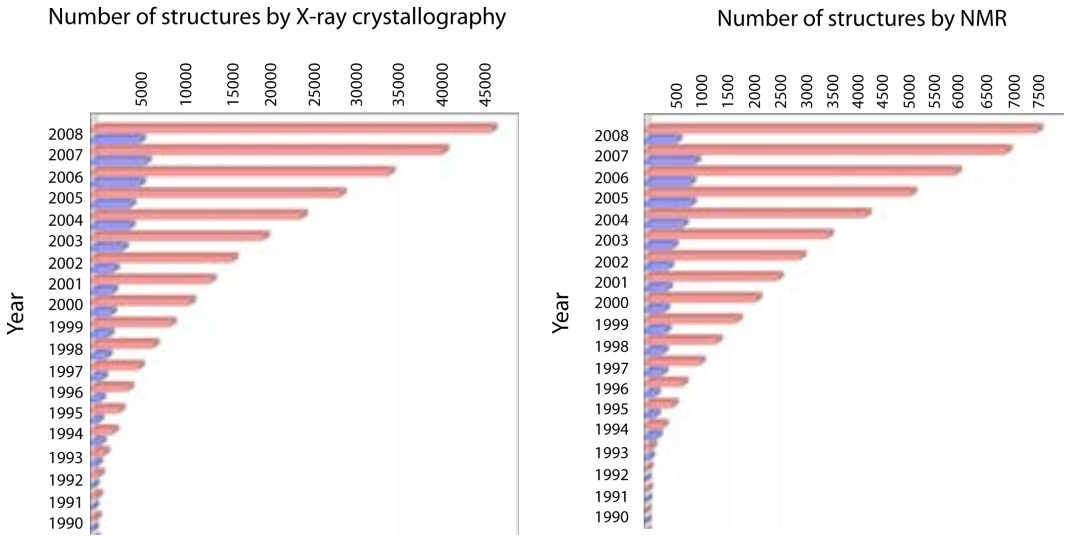



**Figure 4: Year growth of structures solved by X-ray crystallography and NMR**.

Statistics have been obtained from the Protein Data Bank (PDB )(http://www.rcsb.org/pdb).

### 1.3.1 Preparation of NMR samples

The protein structure determination by NMR methods requires a purified and highly concentrated protein sample (typically ~ 0.5-1 mM in solution). In our laboratory, in general, purification is achieved after an affinity chromatography step followed by a gel-filtration chromatography. Usually, for protein structure determination, $^{15}$N- and $^{15}$N/$^{13}$C-labelled samples are required to perform triple-resonance experiments that greatly facilitate the process of data assignment. To obtain these isotopically labelled samples we need to produce the samples in minimal medium, using $^{15}$N- and $^{15}$N/$^{13}$C- as the sole sources of nitrogen and carbon. The protocols for preparing the minimal media are described in Appendix A1. Once we have achieved an optimal protein sample (in terms of purity and concentration) we can start recording the NMR spectra. In Table 1 the NMR experiments utilised in this thesis for the data assignment are listed and in the next sections, a brief description of some of them is provided.

**Table 1:** List of NMR experiments used in protein structure determination. The information obtained from the assignment of the different spectra and the protein labelling required for recording each one are indicated.

| Experiment | | Assignment | Labelling |
|---|---|---|---|
| 1D spectra | $^{1}$H- presat/watergate | protein folding | - |
| 2D spectra | $^{1}$H-$^{1}$H TOCSY | H- side-chains | - |
| | $^{1}$H-$^{1}$H NOESY | NOEs | - |
| | $^{15}$N-$^{1}$H HSQC | $H_N$–N (surface mapping) | $^{15}$N |
| | $^{13}$C-$^{1}$H HSQC | $H_N$–Cx (surface mapping) | $^{13}$C |
| 3D spectra | CBCA(CO)NH | backbone (Cα, Cβ, $H_N$, N) | $^{15}$N, $^{13}$C |
| | CBCANH | backbone (Cα, Cβ, $H_N$, N) | $^{15}$N, $^{13}$C |
| | (H)CC(CO)NH | C- side-chains | $^{15}$N, $^{13}$C |
| | H(CCCO)NH | H- side-chains | $^{15}$N, $^{13}$C |
| | HCCH-TOCSY | H, C- side-chains | $^{13}$C |
| | $^{15}$N-NOESY | NOEs | $^{15}$N |
| | $^{13}$C-NOESY | NOEs | $^{13}$C |
| | HNHA | $^{3}$J ($H_N$, Hα) couplings | $^{15}$N |

### 1.3.2 1D $^{1}$H spectra

The monodimensional $^{1}$H spectrum is the simplest experiment in NMR and can be acquired in few seconds.

For proteins, the vast number of signals are grouped in different regions of the spectrum according to the type of $^{1}$H proton. So, roughly, we find amide protons ($H_N$) at 7-9.5 ppms, alpha protons ($H_\alpha$) around 3.5-5 ppms, methyl groups ($CH_3$) around 0.5-1.5 ppms or aromatic protons around 5.5-7.5 ppms. A clearly distinguishable signal in a $^{1}$H protein spectrum correspond to the indol protons of the tryptophan side-chain, that appear around 10-12 ppms.

But, the most relevant information that we get from a $^{1}$H protein spectrum is that it indicates if the protein is properly folded. As illustrated in Fig. 5 the signals in a folded protein appear widely dispersed in the $^{1}$H spectrum, reflecting a fixed position in a tertiary structure arrangement. In contrast, in an unfolded protein, the lack of structure is reflected in poorly dispersed signals, since the chemical shift of a given $^{1}$H is not perturbed by contacts with the surrounding groups.

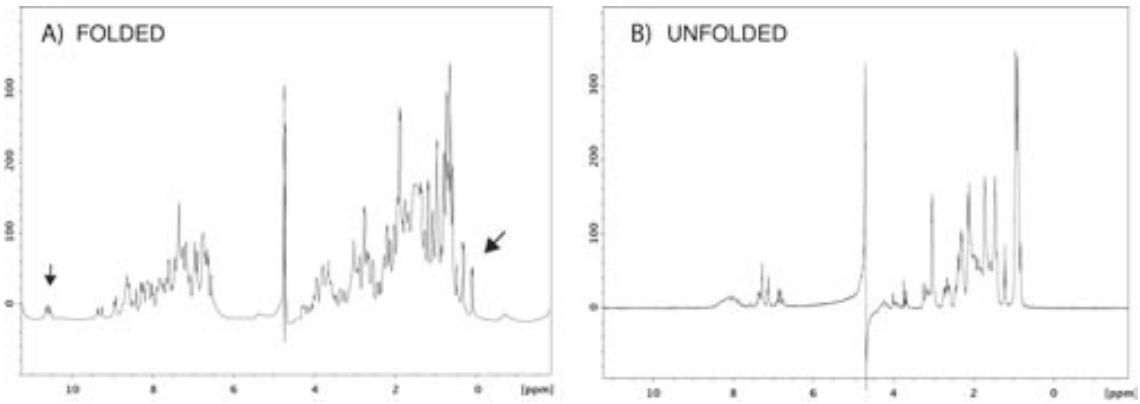



**Figure 5: $^{1}$H spectra of a folded (A) and an unfolded (B) protein.** It is notable the difference in signal dispersion, and also in peak sharpness between the two spectra. Left arrow in (A) indicate the indol peak of tryptophan residues and right arrow mark the better dispersion in methyl region (with shifts below 0 ppm) of spectrum (A) compared to (B).

### 1.3.3 2D $^{1}$H-$^{15}$N HSQC

$^{1}$H-$^{15}$N Heteronuclear Single Quantum Coherence (HSQC) experiments give a proton-nitrogen correlation of all protons attached to nitrogen atoms in the protein. In proteins, the majority of these correlations belong to the amide groups of the peptide bonds. Consequently, as we have one amide proton for every peptide bond (except for those including proline residues), and the number of peptide bonds is the same as the number of amino acids in the protein, in the $^{1}$H-$^{15}$N

HSQC we have, roughly, one peak corresponding to each amino acid of the protein. However, this is only partially true, since in the HSQC we also observe peaks for the side-chains (that contain H-N correlations) of some amino acids, i.e., the side-chains of Asn, Gln, Arg or the indol group of Trp. Pro residues are not observed $^{1}$H-$^{15}$N HSQC, since they do not have amide protons in the peptide bond.

Similarly to $^{1}$H monodimensional experiments and more clearly, the dispersion of the peaks in the $^{1}$H-$^{15}$N HSQC indicates if the protein is folded. In addition, the $^{1}$H-$^{15}$N HSQC also allows us to know if the protein is somehow truncated or degraded (as we would observe less peaks than expected from the number of residues) or if it displays more than one conformation (as we would observe more peaks than expected from the number of residues). Fig. 6 displays the appearance of the $^{1}$H-$^{15}$N HSQC of a folded and an unfolded protein.

One of the most common applications of the $^{1}$H-$^{15}$N HSQC experiment is in the field of protein-ligand interactions, because we can monitor amide proton resonances for every residue of a $^{15}$N-labelled protein. In this way, we can detect if a protein binds to a ligand because amide proton chemical shifts experience changes upon interaction. Moreover, we can map the region of the protein involved in the binding. However, to do this we need to have the peaks of the $^{1}$H-$^{15}$N HSQC identified, and that implies the assignment of the 3D backbone experiments, discussed in the next section.

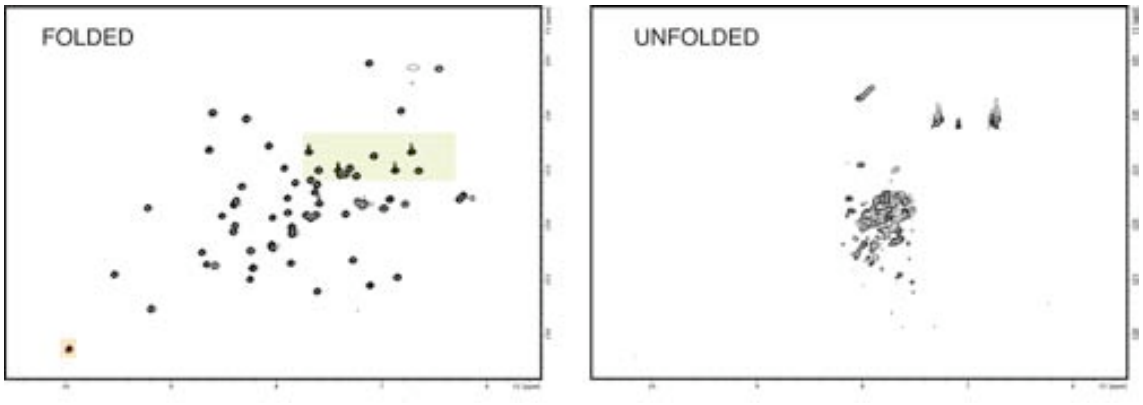



**Figure 6: Picture of a $^{1}$H-$^{15}$N HSQC spectra.**

The horizontal axis corresponds to $^{1}H_{N}$ shifts (around 7-9.5 ppms) and the vertical axis to $^{15}$N shifts (around 100-135 ppms). *Left* spectrum shows the image of a folded protein with widely dispersed signals. The green and orange boxes indicate the typical position for Asn/Gln side-chains and for the Trp indol proton, respectively. *Right* spectrum shows a totally unfolded protein, with amide resonances concentrated in the middle, indicating random coil conformation.

Apart from interaction studies, $^{1}$H-$^{15}$N HSQC are also useful to monitor protein dynamics upon variations of pH, temperature or other parameters.

### 1.3.4 3D CBCA(CO)NH / CBCANH

The first step in the protein structure determination by NMR is the assignment of the CBCA(CO)NH/CBCANH pair of spectra. These spectra are also called 'backbone experiments' since we get the chemical shift information of N, $H_N$, Cα and Cβ, the atoms that conform the polypeptidic chain of the protein.

The strategy used in the work with these spectra is based on the fact that $^{13}C$ chemical shifts of Cα and Cβ are characteristic for each amino acid and are rather insensitive to the tertiary structure. Therefore, their values do not deviate significantly from the standard values, which are listed in Table 2.

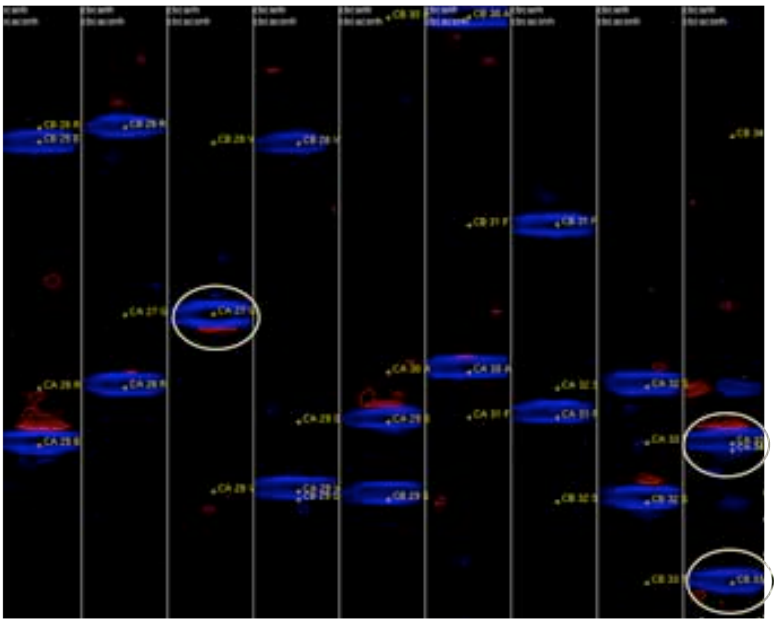
Random coil **C** chemical shifts (ppm)

| | α | β | γ | δ | ε |
|---|---|---|---|---|---|
| G | 44 | | | | |
| A | 52 | 18 | | | |
| D | 53 | 40 | | | |
| E | 55 | 30 | 36 | | |
| S | 58 | 62 | | | |
| T | 60 | 66 | 15 | | |
| C | 56 | 27 | | | |
| M | 54 | 32 | 30 | | 15 |
| V | 60 | 33 | 15 | | |
| L | 54 | 41 | 24 | 20 | |
| I | 59 | 38 | 30 (γ1)<br>12 (γ2) | 8 | |
| F | 57 | 39 | | | |
| Y | 56 | 37 | | | |
| W | 56 | 29 | | | |
| H | 55 | 28 | | | |
| P | 62 | 31 | 26 | 50 | |
| N | 53 | 37 | | | |
| Q | 55 | 25 | 30 | | |
| R | 56 | 30 | 28 | 42 | |
| K | 55 | 31 | 24 | 28 | 40 |

**Table 2: Averaged values of the carbon chemical shifts for the 20 natural amino acids.** Amino acids forming part of α-helical regions have, generally, the Cα and Cβ values shifted to higher and lower ppms respectively. In contrast, in β-sheet regions, Cα are shifted to lower ppms and Cβ are shifted to higher ppms. Thus, backbone experiments are useful to identify secondary structure regions in a protein.

In the CBCA(CO)NH experiment we observe the Cα and Cβ peaks of a *i*-1 residue from the amide N and $H_N$ of a residue *i* while in the CBCANH we observe, from the same residue *i*, the peaks corresponding to Cα and Cβ of either the *i*-1 residue and the own *i* residue. So, as exemplified in Fig. 7, by examining the two spectra simultaneously, we can walk across the sequence of the protein to obtain the backbone resonances for all residues.

**CBCA(CO)NH**



**CBCANH**

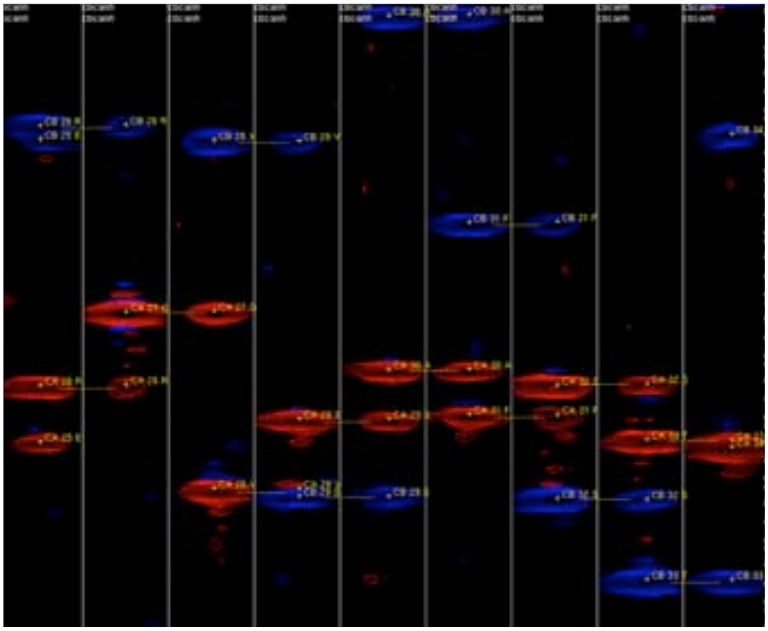


**Figure 7: Display of CBCA(CO)NH and CBCANH spectra.**
The x and y axes correspond to the $H_N$ and C frequencies respectively, and the columns, called *strips*, represent N planes (z-axis), each one belonging to the N resonance of the different amino acids.
CBCA(CO)NH shows two peaks (the Cα and Cβ of the *i*-1 residue) in every *strip*, whereas CBCANH contains, in the same *strip*, the same two peaks and the Cα and Cβ of the next residue (the *i* residue). These two extra peaks in CBCANH appear (at the same C resonances) as the only two peaks in the CBCA(CO)NH of another *strip* (for instance, the *i*+1). In this way, searching through the strips we can correlate all the residues of the protein sequence (as indicated by yellow lines in CBCANH).

It is important to remark that firstly, only the characteristic shifts of the carbon resonances for a given amino acid allow the identification of an amino acid pair among all the strips. Gly, that only have Cα and thus one single peak in the CBCA(CO)NH or Ser and Thr, that have Cα and Cβ peaks inverted at high ppm shifts are examples of these characteristic shifts and are circled in the CBCA(CO)NH spectrum.

However, the complexity in the interpretation and assignment of the backbone experiments increases with protein size due to the overlapping of peaks from different residues.

### 1.3.5 3D side-chain experiments

Once we have completed the backbone resonance assignment, we continue with the assignment of the side-chains, and that includes all the remaining carbon (C) and proton (H) resonances from the different protein residues.

(H)CC(CO)NH allows the assignment of all C resonances and it functions in an analogous manner to backbone experiments, but in this case, from a *strip* belonging to a residue *i*, we observe a variable number of peaks depending on the type of amino acid corresponding to *i*-1. This means, for instance, that if *i*-1 is a Gly we would observe only one peak, but if it is a Lys we would observe up to five peaks (see Table 2).

In H(CCCO)NH we do not look at C resonances but in the correlation between N, $H_N$ and the rest of H resonances, from $H\alpha$ to the H atoms of methyl groups. Thus, the y-axis in this case correspond to H frequencies instead of C frequencies, but the method for the assignment is the same as before, with *strips* that contain a certain number of peaks depending on the concrete amino acid. Furthermore, as we already know N and $H_N$ values from backbone experiments, assignment of peaks in this spectrum is a simple task.

Nevertheless, and unlike C resonances, H resonances are very sensitive to the environment, and thus to protein structure. This means that the chemical shift for a given H resonance can be significantly different from the standard value of the reference tables. Consequently, we can not know the H atom corresponding to a given a peak in a *strip* of the H(CCCO)NH based on the chemical shift. To assign univocally the H atoms we need a direct correlation between the C and H chemical shifts, and we obtain this correlation from the HCCH-TOCSY experiment.

Unlike the 3D experiments analysed until now, we do not generate *strips* of the residues to work with the HCCH-TOCSY. Instead, we assign the peaks of the spectrum directly from the planes, which in this case belong to C resonances and cover the range from $C\alpha$ to C resonances of methyl groups (about 70 ppms, see the values of Table 2). In the x and y axes we have H chemical shifts. Thus, for instance, if we are in a plane that corresponds to a $C\alpha$, the peak of the diagonal should be the $H\alpha$ attached to it, and the related peaks correspond to the other H groups of the residue observed from this $H\alpha$. To clarify this, and example of this assignment process is illustrated in Fig. 8.

Generally, the HCCH-TOCSY is recorded with a sample where the water has been exchanged to deuterium ($D_2O$) in order to improve the peak signals and to avoid the appearance of experimental artefacts.

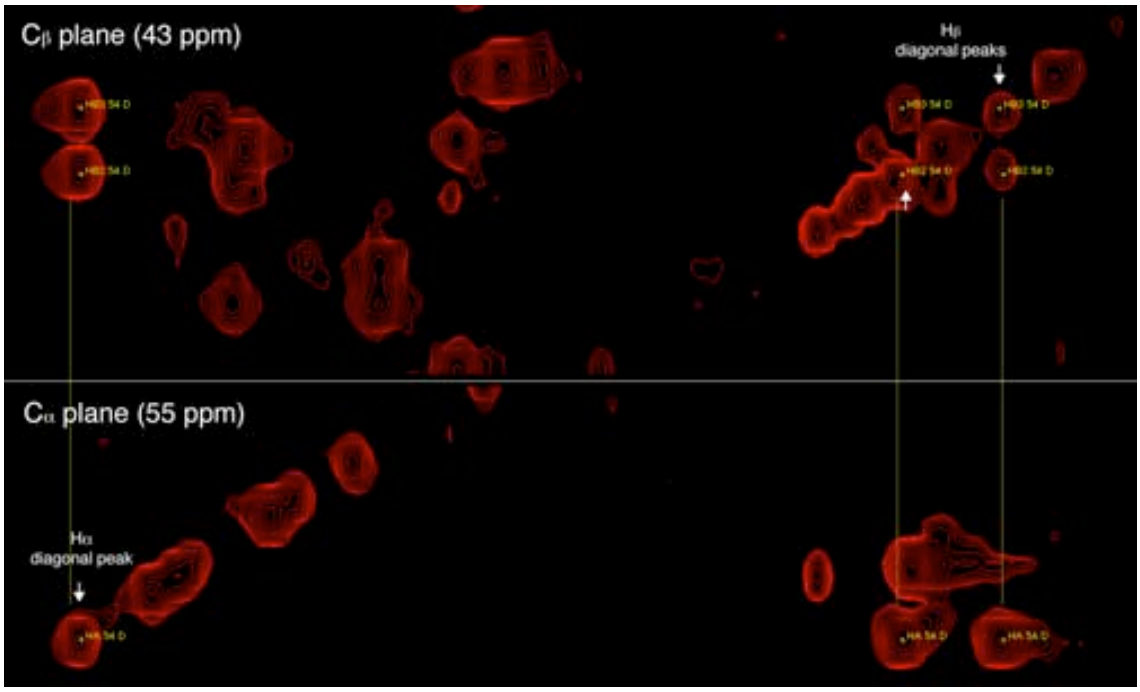



**Figure 8: Assignment process in a HCCH-TOCSY spectrum.**

The picture shows the assignment of a Asp residue, which contains two C (Cα and Cβ) and three H (one Hα and two Hβ). The planes represent the frequencies of C, that we know from the backbone experiments (or from the (H)CC(CO)NH if we have a residue with more C atoms). In the Cα plane, the diagonal peak (marked by a white arrow), which has the two H frequencies equal, correspond to the Hα, and from it we can also observe the two Hβ peaks. Conversely, in the Cβ plane, we observe the same pattern of peaks, but this time the diagonal peaks correspond to the two Hβ peaks, and from each of them we can observe the other Hβ and the Hα. Asp residue has a relatively simple pattern but as the length of the side-chain of a residue increases, so does the complexity of the patterns. It is important to note that from the H(CCCO)NH, the H resonances of the different residues are already known, and this helps us to localize the peaks in the HCCH-TOCSY. However, the H(CCCO)NH experiment itself does not permit to discriminate among the H peaks, as they are not linked to C atoms.

### 1.3.6 2D $^{1}$H-TOCSY

2D $^{1}$H-TOCSY (Total Correlation Spectroscopy) display correlations of spin systems within individual amino acids on the basis of scalar couplings (refer to section 1.2.3). Generally, unlabelled samples are used to acquire $^{1}$H-TOCSY experiments and thus magnetisation is not transferred through the peptide bonds. Hence, in a $^{1}$H-TOCSY we have a single plane composed of two axes of H frequency, with a diagonal that contains all the H peaks and cross peaks that correspond to the H peaks of the same spin system as the diagonal peak (Fig. 9). In other words, all cross peaks from one diagonal peak in a $^{1}$H-TOCSY belong to the same amino acid.

If we have previously completed all the side-chain assignments with the 3D spectra, $^{1}$H-TOCSY does not give too much additional information. However, in a 2D experiment, the resolution in the H frequency is higher compared to 3D spectra and therefore, $^{1}$H-TOCSY helps in solving possible ambiguities in peak assignment. In addition, $^{1}$H-TOCSY experiments (in H2O or D2O samples) greatly help in the H assignment from aromatic groups of the side-chains of Trp, Phe, Tyr and His amino acids.

### 1.3.7 2D $^{1}$H-NOESY

The data that we obtain from the assignment of the 2D $^{1}$H-NOESY (Nuclear Overhauser Effect Spectroscopy) is the main source of information in protein structure determination by NMR. Also, the work with this spectrum is the most time-consuming step in the process of peak assignment, due to the high number of cross-peaks or NOEs (Fig. 9). These NOEs, as explained in section 1.2.4, arise from the spatial correlation between close atoms (in this case, protons) in the tertiary structure.

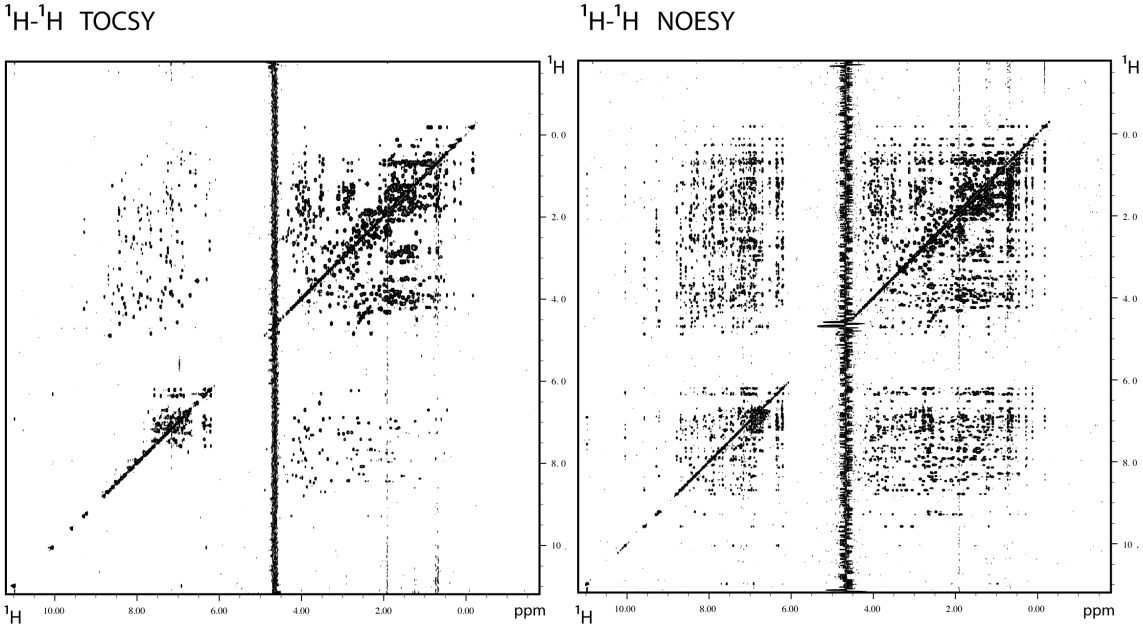



**Figure 9: Images of 2D-homonuclear ($^{1}$H-$^{1}$H) spectra.**

$^{1}$H-TOCSY (*left*) and $^{1}$H-NOESY (*right*) spectra. We clearly observe the difference in the quantity of peaks between the two experiments. Cross-peaks appear repeated at both sides of a given diagonal peak, but only assignment in one side is required for subsequent structure calculation.

In a 2D $^{1}$H-NOESY, from a given diagonal peak, all the related peaks correspond to H atoms that are in spatial proximity to it. This includes of course the H atoms of the same spin system, but also H atoms belonging to other amino acids, which could be far away in the primary

sequence but are near in the folded protein. In fact, these contacts between non-adjacent residues are the most useful ones for the protein structure determination.

In proteins, for instance, H atoms of a residue buried in the hydrophobic core should have many NOEs compared to those of a residue present in an exposed, flexible loop region.

Also, the intensity (or volume) of a NOE is a measure of how far are the two H atoms implicated.

3D $^{15}$N-NOESY and $^{13}$C-NOESY can also be utilised to assign NOE peaks and are complementary to the assignment of 2D $^{1}$H-NOESY. As the size of the protein increases, the work with 3D NOESYs could facilitate the assignment of NOEs, since 2D $^{1}$H-NOESY become too crowded spectra due to the very high number of residues and cross-peaks.

### 1.3.8 3D HNHA

In this spectrum, we observe a correlation between the N, the $H_N$ and the Hα atoms. The importance of this spectrum is that from the relative peak intensity of the Hα respect to the $H_N$, we can extract values of $^{3}$J ($H_N$, Hα) and so, information about the dihedral angle Φ between these two atoms in the polypeptidic chain. In addition, we have indirect information of the secondary structure regions in the protein, because, roughly, Φ values are smaller in α-helices and bigger in β-sheets and loop regions.

The data about dihedral angles extracted from the HNHA is also introduced as restraints, in addition to NOEs, for the subsequent structure calculation.

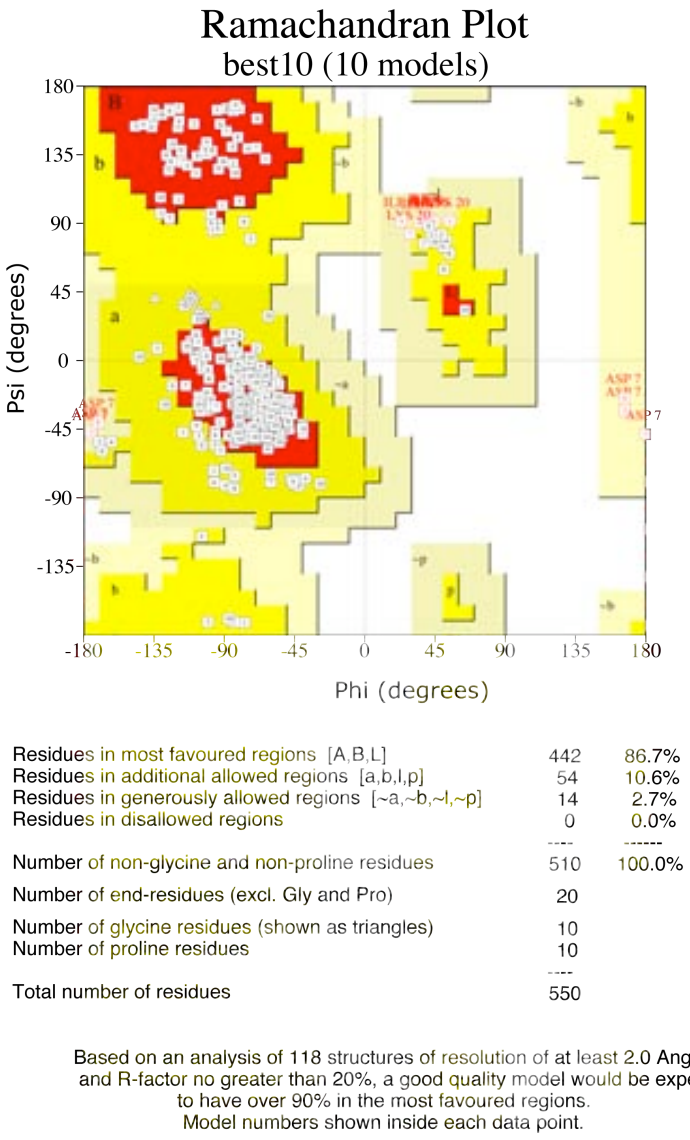
### 1.3.9 Structure determination

For the structure calculation, distance restraints derived from the integration of peaks from NOESY experiments are used, together with dihedral angle restraints. Other sources of structural information that could be used are restraints derived from hydrogen bonds or the measurement of residual dipolar couplings (RDCs), that give information of the relative orientation between two protein regions that are far apart in the structure.

All this information is the input for the structure calculation programmes. The calculation protocols used in our laboratory are based on CNS (Brünger et al, 1998) and calculate an ensemble of structures after an iterative process that tries to minimize the energy of the generated conformers satisfying the maximum number of restraints introduced.

Analyses and comparisons between the final ensemble of conformers dictates the quality of the overall protein structure obtained. For the analysis, we look at the RMSD (Root Mean Square Deviation), a parameter that assesses how good is the convergence between the calculated structures. Another tool for the analysis is the Ramachandran plot, used to check the agreement

of the dihedral angles Φ and Ψ obtained from the calculated structures against the Φ and Ψ allowed conformations in polypeptides (Fig. 10).




| | | |
|---|---|---|
| Residues in most favoured regions [A,B,L] | 442 | 86.7% |
| Residues in additional allowed regions [a,b,l,p] | 54 | 10.6% |
| Residues in generously allowed regions [~a,~b,~l,~p] | 14 | 2.7% |
| Residues in disallowed regions | 0 | 0.0% |
| Number of non-glycine and non-proline residues | 510 | 100.0% |
| Number of end-residues (excl. Gly and Pro) | 20 | |
| Number of glycine residues (shown as triangles) | 10 | |
| Number of proline residues | 10 | |
| Total number of residues | 550 | |

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.
Model numbers shown inside each data point.

**Figure 10: Ramachandran plot**
Picture of a Ramachandran plot for an ensemble of 10 protein structures. We observe that the majority of the residues fall in the most favoured regions (showed in red) for the Φ and Ψ combinations. In fact, this high % of residues in the most favoured regions and the little occurrence of residues in the generously or disallowed regions is indicative of a good quality structures.
Moreover, the higher proportion of residues in values of Φ and Ψ around -45º is indicative of α-helix predominance for this protein. Conversely, residues in the region around 90º, -135º are indicative of β-sheet secondary structure.

# PART 2 - PROTEIN DOMAINS

A protein domain is defined as a part of a protein that can fold autonomously or, in other words, as the fundamental unit of tertiary structure. Proteins can be composed of either one or multiple domains. In general, different domains in a protein are associated with different functions.

Domains are formed by different combinations of secondary structural elements and motifs, that is, α helices and β strands connected by loop regions. In this way, domains can be classified in three main groups: α domains, β domains and α/β domains.

Other possible classifications of protein domains are based on the sequence similarity, the sub-cellular location or the biological function.

## 2.1 Protein interaction domains

Cellular processes require protein domains to direct interactions to other polypeptides, phospholipids, nucleic acids or small molecules. In this way, they participate in and regulate essential biological functions including cell growth, differentiation, cell polarity or apoptosis.

Protein interaction domains mediate associations between proteins in the cell. They are, generally, 30-200 amino acids in size and there are more than 60 protein interaction domain families described. Frequently, protein interaction domains are present in multiple copies in a single polypeptide chain.

Protein interaction domains bind exposed sequences on their protein ligand partners. Typically, they recognize a sequence determinant with other flanking residues that provide additional contacts and increased selectivity to the targets. For example, SH3 domains bind ligands that contain polyproline motifs with the consensus sequence PXXP (with X being any amino acid) and they are classified into two classes according to the orientation in which the motif is positioned with respect to the domain in the bound state. Furthermore, ligand orientation depends on the residues surrounding the PXXP region (Feng et al, 1994).

The consensus sequences recognized by protein interaction domains have been used to predict putative binding partners for a protein containing these domains. However, this is a limited approximation since protein interaction domains are highly versatile in their binding properties. Again, in SH3 domains, apart from the classical binding to the PXXP motif, a subset of these domains can bind the RXXK motif (Liu et al, 2003). Similarly, the second WW domain of CA150 has been shown to accommodate two types of motifs, PPXPP and PPXXPP (Ramirez-Espain et al, 2007).

Furthermore, not all protein interaction domains recognize short peptide motifs. Instead, a number of domains mediate domain-domain interactions and certain domains can have the two types of binding. This is the case for the PDZ domain, which in general binds short peptide

motifs (~ 4 amino acids) at the C-terminal of their ligand partners, but can also mediate PDZ-PDZ domain interactions (Hillier et al, 1999).

A number of protein interaction domains are specialized in detecting post-translational modifications. A classical example is the SH2 domain, which strictly interacts with phosphotyrosine-containing peptides. Similarly, FHA domains recognize epitopes that contain a phosphothreonine.

Finally, some protein interaction domains are built from small repeated motifs (up to ~ 50 repeats) to generate a higher-order structure with multiple binding properties. HEAT and TPR domains, among others, are included in this category. For instance, TPR motifs consist in 3-16 tandem repeats of about 34 amino acids, with a helix-turn-helix arrangement and mediate protein-protein interactions as well as the assembly of protein complexes. They participate in various processes such as cell cycle regulation, protein transport or transcriptional control (D'Andrea & Regan, 2003).

An example of different classes of protein domains is illustrated in Fig. 11.

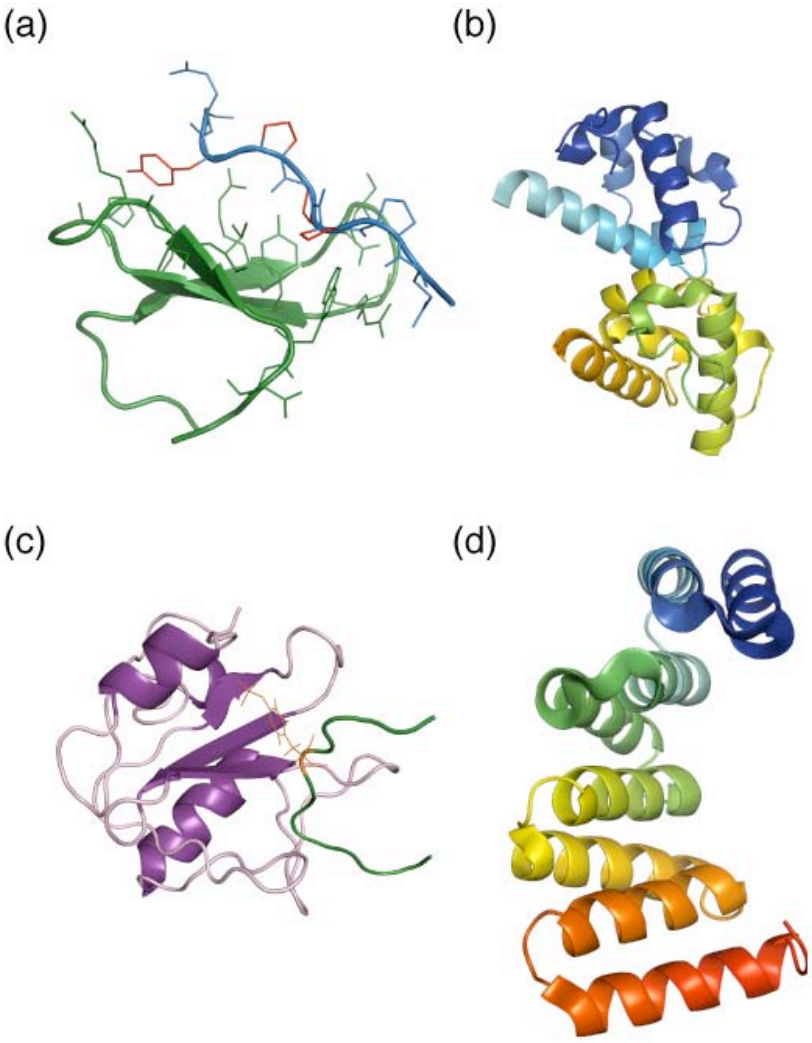


**Figure 11: Classification of protein interaction domains.**
(A) Itch WW3 domain in complex with Epstein-Barr LMP2A derived peptide. The PPXY motif is marked in red (Morales et al, 2007) (B) Yan SAM/Mae SAM domain complex (Qiao et al, 2004) (C) SHC SH2 domain in complex with a Tyr phosphorylated peptide from the T-cell receptor (Zhou et al, 1995) (D) Crystal structure of a 8 repeat consensus TPR superhelix (Kajander et al, 2007)

## 2.2 FF domains

The FF domain is a protein interaction domain reported by the first time in 1999 (Bedford & Leder, 1999) and described as a motif that often accompanies WW domains, as it was identified in human FBP11 and CA150, two proteins that contain WW domains. The domain is, generally, ~ 55 residues in length and was named FF because it harbours two highly conserved phenylalanine residues in its sequence. FF domains have a rather limited distribution in proteins among eukaryotes. Typically, the number of repeated FF domains in proteins ranges between two and six, either consecutively arrayed or spaced by linker regions.
The first structure reported (Allen et al, 2002) showed that FF is an all-helical domain that displays a $\alpha 1$- $\alpha 2$-$3_{10}$- $\alpha 3$ topology (Fig. 12).

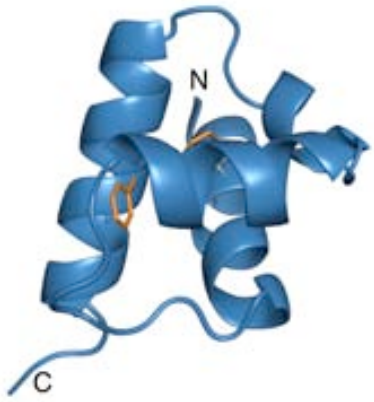



**Figure 12: Solution structure of the first FF domain from HYPA/FBP11.** The two conserved Phe that give name to the domain are depicted in orange.

Since then, there has been only another FF structure reported (Gasch et al, 2006), although there is several FF domain structures deposited in the Protein Data Bank (PDB) (Berman et al, 2000) by the RIKEN Structural Genomics/Proteomics Initiative (RSGI). Thus, compared to other protein interaction domains, there is still little structural information available for the FF domains.

The first specific function described for tan FF domain was the binding to the C-terminal repeat domain (CTD) of RNA polymerase II (Carty et al, 2000) (Appendix A2). Moreover, the interaction only occurred with the CTD in a phosphorylated state. Hence, the FF domain was described as a phosphoserine-binding domain, a definition maintained to date.

Nevertheless, FF domains have been reported to bind other factors that do not contain a phosphoserine target, i.e., the interaction of the Prp40 FF1 domain to a TPR motif of the Clf1 splicing factor (Gasch et al, 2006). In addition, the distinct nature of this interaction reveals that FF domains seem to have a high variability regarding their binding properties.

FF domains are distributed in three main groups of proteins: two groups include nuclear proteins related to splicing (HYPA/FBP11, Prp40 and URN1), and transcription (CA150) and the third group is formed by the cytoplasmatic p190-RhoGAP family of RhoGTPase-activating proteins. In the next sections, an overview on the function and cellular role of the proteins whose FF domains have been part of this study is presented.

### 2.2.1 Prp40 and URN1 splicing factors

The *Saccharomyces cerevisiae* proteins Prp40 and URN1 participate in the splicing process. Briefly, the splicing is the mechanism by which the primary RNA product that results from DNA transcription is processed to generate the mature mRNA that will be further translated into a functional protein. Concretely, the introns, the stretches of sequences that are not translated to proteins, are eliminated from the pre-mRNA and the exons, the coding regions, are joined. Splicing occurs via two transesterification reactions, illustrated in Fig. 13.

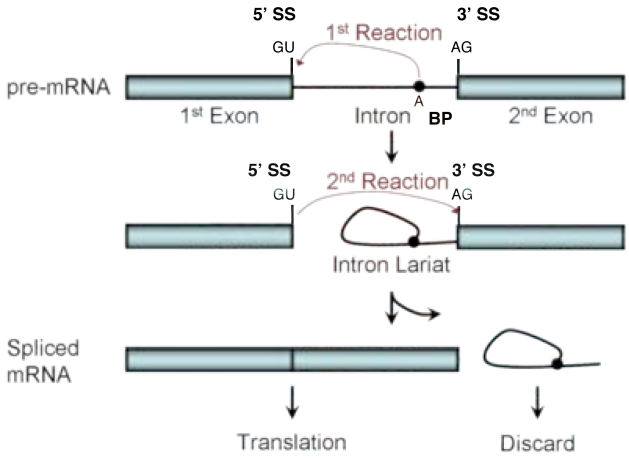



**Figure 13: Mechanism of splicing.**

In the first reaction, there is a nucleophilic attack of the BP adenosine via its 2'OH group to the 5'SS. In the second reaction, the free 3'OH from the 5'ss attacks the 3'ss to join the exons, and the intron – in a

closed conformation termed lariat – is released and discarded.

Introns can vary in length and sequence, but they contain three conserved regions: the 5' and the 3' splicing sites (ss), consisting in the dinucleotides GU and AG respectively, and one adenosine (A) at the branch-point (BP). These conserved regions are indispensable, as they participate in the transesterification reactions and are specifically recognized by several factors during the splicing process.

The splicing mechanism is catalyzed by the spliceosome, a molecular machinery, conserved from yeast to humans, composed by five small nuclear ribonucleoproteins (snRNPs), named U1, U2, U4, U5 and U6, and a wide number of accessory proteins termed splicing factors.

Prp40 is one of these splicing factors, identified as a protein tightly associated with the U1 snRNP (Kao & Siliciano, 1996). It is composed by a pair of WW domains in the N-terminal followed by four FF domains (Appendix A3). Prp40 was proposed to be homologous to human HYPA/FBP11, as they display a similar domain organization and have related interactions (Bedford et al, 1998). However, a more recent study performed in our group showed that Prp40 and HYPA/FBP11 are not orthologous, since not all their FF domains cluster together in phylogenetic analysis, as some Prp40 FF domains are more closely related to FF domains of human CA150 transcription factor (Gasch et al, 2006).

The Prp40 WW domains are implicated in cross-intron bridging via a direct interaction with the branch-binding point (BBP) protein, in the early steps of splicing (Abovich & Rosbach, 1997) (Fig. 14).

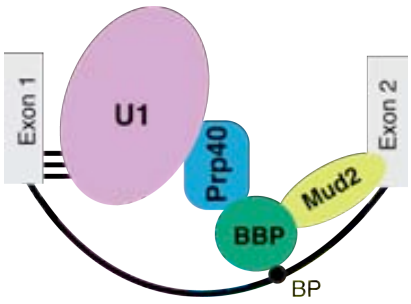



**Figure 14**: **Prp40 brings the 5' ss and the BP into spatial proximity.**

Prp40 also associate with Prp8, a U5 snRNP core component, capable of interacting directly with the 5'ss, the 3'ss and the BP (Teigelkamp et al, 1995) (MacMillan et al, 1994), that acts as a platform to hold various components (reviewed in (Grainger & Beggs, 2005)). Initially, the interaction was believed to involve the WW domains of Prp40 and a PPXY motif in the N-terminal of Prp8 (Abovich & Rosbach, 1997). However, a later study showed that Prp40 WW

domains possess a greater preference for PPΨΨP motifs (Ψ being an aliphatic residue) and only interacted weakly with PPXY motifs (Wiesner et al, 2002).

The observation that the Prp40 region comprising FF domains interacted with the Clf1 (crooked neck like factor), a splicing factor required for the 5'ss pre-mRNA cleavage (Chung et al, 1999), led to a detailed mapping of this binding, that implicated the Prp40FF1 domain and a N-terminal TPR motif of the Clf1 (Gasch et al, 2006).

Prp40 has also been implicated in interactions with the phospho-CTD at multiple locations, including the WW and FF domains (Morris & Greenleaf, 2000). After, though, it was proved that neither the WW domains of Prp40 nor the FF1 possessed ability to bind the phospho-CTD (Gasch et al, 2006; Wiesner et al, 2002).

Contrarily to the extensive information existent on Prp40, the URN1 protein (formerly known as Ypr152) is almost completely unknown. URN1 is a nuclear protein that may have a role in splicing, since its architecture resembles that of Prp40, with one WW and one FF domain. In fact, Prp40 and URN1 are the only FF-containing proteins in *Saccharomyces cerevisiae.*

In the Saccharomyces Genome Database (SGD) (Hong EL et al) URN1 is reported as a putative pre-mRNA splicing factor associated with the U2-U5-U6 snRNPs and the splicing complexes RES (Dziembowski et al, 2004) and the Prp19-associated (NTC) (Chan et al, 2003; Chen et al, 2002), but no published data is available.

Only, in the same study performed with Prp40, it was observed that URN1 could not bind the phospho-CTD (Morris & Greenleaf, 2000).

### 2.2.2 CA150 transcription factor

Eukaroryotic gene expression is a multistep process that begins with transcription by RNA polymerase II. During transcription, the pre-mRNA transcript undergoes several processing steps including capping at the 5' end, splicing of introns and polyadenilation at the 3' end. These processes are followed by other post-transcriptional events such as mRNA export, translation and degradation. Recent studies have shown that there is a molecular cross-talk of the individual steps, which means that some components participate in more than one of the events involved in gene expression (reviewed in (Hagiwara & Nojima, 2007))

One of these components is the human transcription elongation repressor CA150 (also known as TCERG1), which has been involved in the coupling of transcription and splicing processes (Sanchez-Alvarez et al, 2006; Smith et al, 2004; Sun et al, 2004), although its precise function has not been elucidated to date.

Like Prp40 and URN1, CA150 has a modular architecture consisting in N-terminal WW domains followed by FF domains (Appendix A3).

CA150 was originally identified as a protein associated with the RNA polymerase II holoenzyme with a role in the regulation of Tat-mediated transcription from the human immunodeficiency virus type 1 (HIV-1) (Sune et al, 1997). Additional experiments proved that CA150 acts as an inhibitor of transcription elongation in a promoter-specific manner (Sune & Garcia-Blanco, 1999) and a recent work defined CA150 to have a central function in mRNA processing (Pearson et al, 2008).

Studies on CA150 associations led to the findings that the WW domains are responsible for the binding to SF1 (Goldstrohm et al, 2001). This splicing factor – the human orthologous of yeast BBP protein- have been shown to play a role in transcription repression, (Zhang & Childs, 1998; Zhang et al, 1998) and therefore, it is likely that CA150 mediates this function.

On the other hand, it has been described that the CA150 FF domains interact specifically with the phospho-CTD repeats (Carty et al, 2000). Besides, another work identified multiple transcription- and splicing-related factors as interacting partners for the FF domains, and their detailed study of the binding to Tat-Sf1 derived motifs suggested a binding mode for the FF domains to their targets through multiple weak interactions in a non-cooperative manner (Smith et al, 2004). More recently, FF domains have been found to be essential for co-localization of CA150 to speckles, sub-nuclear structures enriched in pre-mRNA splicing factors. The same work suggested the possibility that CA150 could serve as a carrier of splicing and other factors from the speckles to transcription sites (Sanchez-Alvarez et al, 2006).

### 2.2.3 p190-A and p190-B RhoGAPs

RhoGTPases are a family of signalling G proteins included in the Ras superfamily of GTP hydrolases. Specifically, RhoGTPases are small proteins (~ 20 kDa) found in all eukaryotes from yeast to mammals, with a widely studied role in the regulation of actin cytoskeleton dynamics, even though they are involved in the control of many other signal transduction pathways in the cell. RhoGTPases can be grouped in three main subclasses: Cdc42, Rac and Rho proteins. In mammals, RhoGTPases comprise a family of 20 members.

Generally, they act on a specific process through interactions with downstream effector proteins. Some of them are listed in Table 3, extracted from (Bustelo et al, 2007).

RhoGTPases have been defined as molecular switches due to their fluctuation between an active GTP-bound state and an inactive GDP-bound state (Fig. 15). The cycling between these two forms is regulated by three groups of proteins: the GEFs (guanine nucleotide-exchange factors), the GAPs (GTPase-activating proteins) and the GDIs (guanidine nucleotide-dissociation inhibitors) (reviewed in (Etienne-Manneville & Hall, 2002)).

| Rho GTPase | effector protein | biological process |
|---|---|---|
| RhoA | Itpr1 (Inositol 1,4,5-triphosphate receptor) | Calcium entry in endothelial cells |
| | PlcG1 (Phospholipase, C type) | Production of second messengers |
| | KcnA2 (Potassium Channel subunit) | Potassium entry |
| | Ppp1r12A (Regulatory subunit of phosphatase 1) | Myosin light chain inactivation |
| Cdc42 | Mig-6 (Scaffold protein) | Activation of the Jnk route |
| | Tnk2 (Tyrosine kinase) | Signal transduction, activation of GEFs |
| | CopG2 (Coatomer protein) | Vesicle trafficking (clathrin route) |
| | Map3K10 (Serine/threonine kinase) | Activation of kinase cascades |
| Rac1 | SynJ2 (Polyphosphoinositide phosphatase) | Inhibition of receptor endocytosis |
| | CybA (NADPH oxidase complex subunit) | Superoxide production |
| | Fhod1 (Formin-like) | Cytoskeletal and transcriptional regulation |
| | Nos2A (Nitric oxide synthase) | Nitric oxide production |

**Table 3**: **List of selected effector proteins and cellular processes for RhoA, Cdc42 and Rac1.**

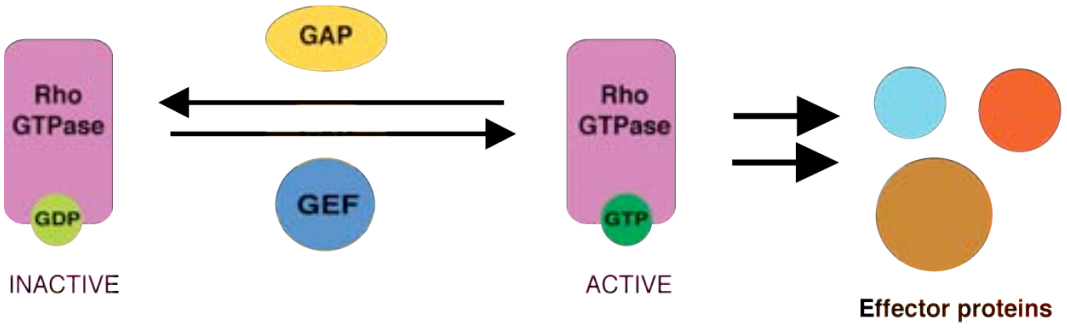



**Figure 15: RhoGTPase activity depends on GDP-GTP exchange.**

GEFs activate RhoGTPases by promoting the exchange of GDP for GTP. On the contrary, GAPs inactivate RhoGTPases by enhancing their intrinsic GTPase activity. GDIs (not shown) attach to the C-terminal of RhoGTPases and stabilize the GDP-bound form.

RhoGAPs greatly increase the intrinsic GTPase activity of RhoGTPases. In this way, they cause their inactivation. Also, RhoGAP activity is tightly regulated by a wide number of mechanisms, including phosphorylation, protein-protein interaction, lipid binding or proteolytic degradation. Around 70 proteins in eukaryotes have been identified that form part of the RhoGAP family. Thus, compared with the number of RhoGTPases they regulate, there is an apparent excess of RhoGAPs. As reviewed in (Tcherkezian & Lamarche-Vane, 2007) there are four possible explanations to this phenomenon:

- RhoGAPs have selective tissue expression and tissue-specific functions

- different RhoGAPs act on the same single RhoGTPase.
- each RhoGAP selectively regulate a specific RhoGTPase signalling pathway.
- the RhoGAP domain might act only as a recognition module to facilitate the cross-talk between RhoGTPase-mediated pathways and other signalling pathways in the cell.

This last possibility is linked to the fact that RhoGAP proteins contain, apart from a RhoGAP domain - consisting of around 200 residues with an all-helical architecture - different other functional domains involved in a variety of biological processes.
Furthermore, this multi-domain composition of RhoGAP proteins and sequence homology studies have allowed the classification of RhoGAPs in distinct subfamilies, some of them displayed in Fig. 16, extracted from (Tcherkezian & Lamarche-Vane, 2007).

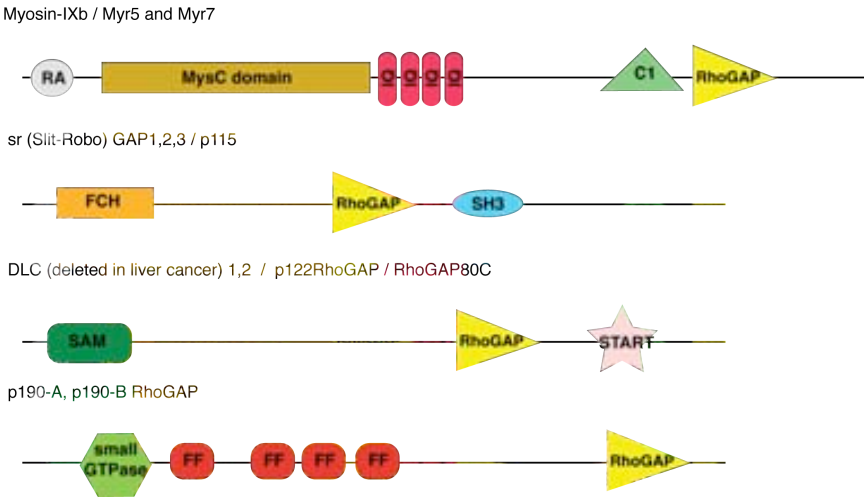



**Figure 16: Examples of RhoGAP subfamilies.**
RhoGAP proteins are formed by a GAP domain an many other functional domains: RA, Ras-association domain; MysC, myosin large ATPases; IQ, calmodulin binding motif; C1 protein kinase C conserved region 1 (cysteine-rich); FCH, Fes/CIP4 homology; SH3, Src-homology 3 domain; SAM, sterile alpha motif; START, STAR (steroidogenic acute regulatory)-related lipid transfer; small GTPase domain of the Ras superfamily; FF domain.

One of these subfamilies is composed by two closely related proteins, p190-A and p190-B. p190 RhoGAPs display four consecutive FF domains (Appendix A3), and are, in fact, the only described cytoplasmatic proteins that contain FF domains.
p190-A and p190-B possess GAP activity towards Cdc42, Rac1 and RhoA *in vitro* (Settleman et al, 1992). However, *in vivo*, p190-A only showed GAP activity towards RhoA to cause inhibition of actin stress fiber formation (Ridley et al, 1993) and to promote cell spreading and migration (Arthur & Burridge, 2001).

p190-A activity is regulated by phosphorylation. For example, the association of p190-A with p120 RasGAP is promoted when p190-A is phosphorylated on residue Y1105, allowing its recognition by the N-terminal SH2 domains of p120 RasGAP (Hu & Settleman, 1997; Roof et al, 1998). The formation of p190 RhoGAP/p120 RasGAP complex stimulates p190-A activity (Bradley et al, 2006) and reduces p120 activity (Moran et al, 1991), and is believed to facilitate cross-talk between Rho and Ras signalling pathways.

Furthermore, p190-A Y1105 residue has been shown to be the target of several kinases to mediate specific functions: c-Src kinase phosphorylates p190-A at this position to facilitate actin stress fiber disassembly in EGF-stimulated fibroblasts (Chang et al, 1995) or, as reported recently, Brk kinase mediated effects on promoting proliferation, migration, invasion and transformation of breast cancer cells involve p190-A phosphorylation on Y1105 (Shen et al, 2008).

But Y1105 is not the only phosphorylation target of p190-A RhoGAP. A study reported that a Tyr residue within the first FF domain of p190-A was a PDGF-α receptor kinase target, and that this mechanism regulated the interaction of p190-A with the TFII-I transcription factor (Jiang et al, 2005). This is the only reported case to date of phosphorylation within a FF domain as a way to modulate the interaction with its binding partner.

# OBJECTIVES

The main subject of this thesis has been the study of the FF domain from a structural point of view using NMR spectroscopy.

The FF domain is a protein interaction domain present, generally, in three protein families: the splicing factors FBP11 and Prp40, the transcription factor CA150 and the RhoGTPase regulatory proteins p190 RhoGAPs. We have divided our work in three chapters, each of them including structural studies on FF domains from the three different families. Briefly, in the first section we have determined the solution structure of two FF domains from the yeast splicing factors Prp40 and URN1 and have performed extensive structural comparisons with all the FF domains available. In the second part, we have studied the interaction of the FF domains from CA150 with the phosphorylated C-terminal domain of the RNA-polymerase II, the first described and most common ligand for the FF domain. Finally, in the third part, we have examined the phosphorylation mechanism of the first FF domain of p190-A RhoGAP, the only reported post-translational modification within an FF domain.

The specific objectives of this thesis are detailed next:

**Chapter 1:** Structural studies on FF domains from yeast splicing factors URN1 and Prp40

1. Determination of the solution structures of URN1FF and Prp40FF4 domains by NMR and comparison to previously described FF structures.

2. Electrostatic analysis of the FF domains surfaces and implications for target binding specificity.

3. Structural similarity searches for the FF domain.

4. Determination of Prp40 FF domains ability to interact with the phospho-CTD.

**Chapter 2:** Binding studies of the CA150 FF domains to phosphorylated-C-terminal domain (CTD) of RNA-polymerase II

1. Determination of CA150 FF domains ability to interact with the phospho-CTD.

2. Characterization of the FF binding surfaces for the phospho-CTD interaction.

3. Comparison of the phospho-CTD binding between individual FF domains and a double construct.

**Chapter 3:** NMR structure of p190-A RhoGAP FF1 and insights into its phosphorylation mechanism

1. Determination of the solution structure of the p190-A RhoGAP FF1 domain.

2. Study of the structural consequences of the phosphorylation on the FF domain.

3. Study of p190-A RhoGAP FF1 and FF4 binding to the N-terminal region of the transcription factor TFII-I.

# CHAPTER 1

STRUCTURAL STUDIES ON FF DOMAINS FROM YEAST URN1 AND PRP40 SPLICING FACTORS

## 4.1 SECTION 1

### 4.1.1 INTRODUCTION

FF domains were first identified as repeat sequences of about 60 amino acids found in the murine splicing factor FBP11 (Formin Binding Protein 11) (Bedford & Leder, 1999). They are present in three protein families: the splicing factors FBP11, Prp40 and URN1, the transcription factors CA150, and the p190RhoGTPase-related proteins (Bedford & Leder, 1999).

The number of FF domains in protein sequences ranges, in general, between two and six. Structures of two FF domains (the first FF domains from FBP11 and Prp40 proteins) have been analyzed in detail (Allen et al, 2002; Gasch et al, 2006). Four CA150 FF domains (FF1, FF2, FF3 and FF4) and one FBP11FF5 structures have also been deposited in the PDB by the RIKEN consortium in 2006. FF sequences are divergent (Gasch et al, 2006; Sun et al, 2004) but the overall fold is conserved, consisting of three alpha helices and a short $3_{10}$ helix, arranged in a $\alpha1$-$\alpha2$-$3_{10}$-$\alpha3$ topology.

In contrast to other well-characterized protein-protein interaction domains such as PH, PDZ, WW or SH3 domains, in which targets were identified and complexes solved soon after the domains were discovered, ligand recognition by FF domains remains an open question (Bhattacharyya et al, 2006). Studies performed to date show that different FF domain constructs can interact with different kinds of ligands, namely the phosphorylated CTD repeats (Carty et al, 2000), the TFII-I transcription factor (Jiang et al, 2005), the $(D/E)_{2/5}$-F/W/Y-$(D/E)_{2/5}$ consensus sequence present in Tat-SF1 (Smith et al, 2004), a tetratricopeptide repeat (TPR) motif from the splicing factor Clf1 (Chung et al, 1999; Vincent et al, 2003; Wang et al, 2003), and even with RNA (Sun et al, 2004).

Unfortunately there is not enough structural information to interpret all these interactions. Systematic comparisons of FF sequences have been previously performed by others and us, with the aim of clarifying domain composition and similarity in related sequences (Gasch et al, 2006; Sun et al, 2004). These analyses however, did not provide patterns of conserved residues in the sequence that could indicate the presence of conserved binding sites. On the contrary, they could only demonstrate that FF domain sequences are divergent, with the most variable part localized around helix 2 and loop 2, which includes the $3_{10}$ helix. Until now, only two binding sites have been characterized at the residue level, the phospho-CTD binding site of the HYPA/FBP11FF1 and that of Prp40FF1 for the TPR motif present in the splicing factor Clf1. In the first case the FF domain binds to the ligand using a positively charged patch. In the case of Prp40 FF1 domain, the interaction involves the DxR(Y/F) motif (semiconserved in the FF family) encircled by negatively charged and neutral/aromatic residues (Allen et al, 2002; Gasch et al, 2006). However, the Prp40 FF1 domain is unable to interact with the phosphorylated

sequence that binds to FBP11FF1. Furthermore, both binding sites are located in different places, demonstrating not only that FF domains can recognize several ligands but also that at least two FF sequences (Prp40FF1 and FBP11FF1) display different binding sites (Gasch et al, 2006).

To gain insight into the molecular function of FF domains, we selected the FF domain present in the URN1 yeast splicing factor since it is one of the two proteins containing only one FF domain. In addition, and like all these other FF-containing splicing factors, URN1 contains a WW domain. URN1 is described in the *Saccharomyces* Genome Database (SGD) database (Hong EL et al) as a pre-mRNA splicing factor associated with the U2-U5-U6 snRNPs, the RES complex, and the Prp19-associated complex (NTC). A number of URN1 binding partners are also reported in interaction databases, but no specific targets have been described for the URN1FF domain.

To examine the FF sequences from a different perspective we set to compare the URN1FF structure to all FF domain structures available, and assayed the possibility of identifying potential binding sites. In this comparison we included an analysis of electrostatic surfaces. We found that although the fold is conserved, the electrostatic distribution is variable, even for domains with similar overall $pK_a$ values*. This suggests that charge distribution on the surface may describe ligand recognition sites better than $pK_a$ values, assuming that binding to charged motifs is partially driven by electrostatic interactions.

To expand our analysis further we also compared the URN1FF structure to other structurally similar proteins, as reported in the Structural Classification Of Proteins (SCOP) database and to SURPs domains since they share with FF domains the same $\alpha1$-$\alpha2$-$3_{10}$-$\alpha3$ architecture (Denhez & Lafyatis, 1994; Kuwasako et al, 2006; Murzin et al, 1995; Spikes et al, 1994). From our structural analysis, we suggest that FF domains are, at least, as related to SURP domains as they are to the structures reported in the SCOP database. Furthermore, it would be interesting to find whether some FF domains may share with these structures not only some characteristics of their binding sites but also their ability to recognize similar ligands.

* The overall $pK_a$ value is used to refer to what is also known as the isoelectric point of the protein (pI). Although the utilisation of the pI is more extended the two nomenclatures are accepted, and all along this work we have used the term overall $pK_a$ value.

### 4.1.2 EXPERIMENTAL PROCEDURES

*Sample Preparation*- URN1FF domain construct corresponds to residues 212-266. The fragment was amplified by PCR using genomic DNA of *Saccharomyces cerevisiae* as template and cloned into a pETM30 vector for the production of a N-terminal fusion protein with a His tag followed by GST and a TEV protease cleavage site. For the protein expression, BL21 (DE3) cells were grown to 0.6 optical density and induced with 0.4 mM IPTG at 20ºC overnight. Purifications were achieved after a glutathione sepharose affinity chromatography, a subsequent cleavage with TEV, a $Ni^{2+}$ chelation step for removal of the GST and the uncleaved fusion protein and a final gel filtration chromatography step on a HiLoad Superdex™ 75 prepgrade (GE healthcare Life Sciences). $^{15}N$ and $^{13}C$ -labelled protein was prepared following the method developed by Marley and co-workers (Marley et al, 2001) using D-[$^{13}C$]-glucose and $^{15}NH_4Cl$ as sole sources of carbon and nitrogen respectively. URN1FF sample for NMR experiments was concentrated to 0.5mM in 20mM sodium phosphate buffer, 130mM NaCl, 0.03% (w/v) $NaN_3$ in 90% $H_2O$, 10% $D_2O$ or 100% $D_2O$ at pH 5.8.

*NMR spectroscopy*- All NMR data were acquired at 285K on either Bruker DRX-600 or Bruker DRX-800 NMR spectrometers. Main-chain and side-chain assignments were obtained combining the information from standard triple resonance experiments (CBCA(CO)NH, HNCBCA, HCCCONH, HCCH-TOCSY) and homonuclear 2D TOCSY (65 ms mixing time) and 2D NOESY (120 ms mixing time) experiments. Intra- molecular proton distance restraints were obtained from peaks assigned in 2D-NOESY, $^{15}N$-NOESY and $^{13}C$-NOESY experiments. All spectra were processed with the NMRPipe/NMRDraw software (Delaglio et al, 1995) and were analyzed with CARA (Bartels et al, 1995). Spectra used for the calculation were integrated with the batch integration method of the XEASY package.

*NMR titration experiments with the TPR repeat*- For the $^{15}N$-HSQC experiments, $^{15}N$-labelled URN1FF domain was prepared at 0.25 mM concentration, in the same buffer as described above, and unlabelled ligand was added to the $^{15}N$-labelled URN1FF up to a final molar ratio 3:1. Measurements were performed at 285K on a Bruker DRX-600.

*Structure Calculation*- For the structure calculation, distance restraints derived from NOESY experiments, $^3J(H^N, H^\alpha)$ obtained from HNHA spectra and hydrogen bond restraints determined by $D_2O$ exchange were used. The structures were calculated using CNS (Brünger et al, 1998) with Structcalc, an *in house* modified protocol of Aria 1.2 (Nilges et al, 1997). Since only unambiguously assigned restraints were used, the protocol was reduced to 2 iterations of 1 and 60 structures respectively, using 100000 cooling steps. All calculated structures were

submitted to water refinement and were ranked based on minimum values of energy-terms and violations. The water refinement protocol was also modified by weighing the value of unambiguous NOEs, hydrogen bonds and dihedral restraints by a factor of ten. In this way all experimental restraints are used during the refinement process and the obtained structures are in better agreement with the experimental data, while retaining good Ramachandran values.
MOLMOL (Koradi et al, 1996) and PyMOL (DeLano, 2002) programs were used to display the result of the structural superimpositions and to generate the figures. Analysis of the quality of the 15 lowest energy structures was performed using PROCHECK-NMR (Laskowski et al, 1996). The statistics from the analysis are shown in Table 4.

*Sequential and structural alignments*- Sequence alignments for the FF and SURP families and profile alignment between them were performed using ClustalX (Thompson, 1997). We selected for the alignments the proteins from each family whose structure has been deposited in the PDB (Berman et al, 2000).
Structural alignments were generated comparing the URN1FF sequence to each of the available FF structures. We employed two strategies, either a global fitting using the CEalign program (Shindyalov & Bourne, 1998) (used as a plug-in in PyMOL), or a manual selection of secondary structure elements. In the first case the program selects a certain number of alpha carbons for the superimposition in order to obtain a best fit of the two structures. In the second case we defined an equal number of residues and used the backbone heavy atoms for the superimposition. In this second case the alignments were performed either with the optAlign mode of the CEalign or directly with MOLMOL.

*Calculation of the electrostatic charge distribution*- The APBS (Adaptative Poisson-Boltzmann Solvent) program was used, as a plug-in in PyMOL, to generate all electrostatic representation of the surfaces (Baker et al, 2001). Several structures were analyzed to determine if the side-chain orientation of charged residues was conserved among the different calculated structures. Since this was the case, only the structure with the lowest energy is displayed in the figures.

*Deposition of assignments and coordinates*- The $^{1}H$, $^{15}N$ and $^{13}C$ chemical shift assignments have been deposited in the BioMagResBank database (http://www.bmrb.wisc.edu) under the accession number 15439, and the protein coordinates in the protein data bank under accession number 2juc (15 structures).

### 4.1.3 RESULTS

*Solution structure of the FF domain in URN1 protein*

Standard multidimensional heteronuclear NMR spectroscopy allowed us to almost completely assign the $^{1}H$, $^{15}N$, and $^{13}C$ resonances. As summarized in Table 4, a total of 1297 non-redundant and unambiguously assigned NOEs together with 56 dihedral angle and 40 hydrogen bond restraints were used for structure determination. A superimposition of the final ensemble of the 15 lowest energy structures (shown in Fig. 17 (a)) is defined with a backbone (N,$C^{\alpha}$,C') RMSD of 0.17 Å. A cartoon representation corresponding to the lowest energy structure showing some representative side-chains is displayed in Fig. 17 (b).

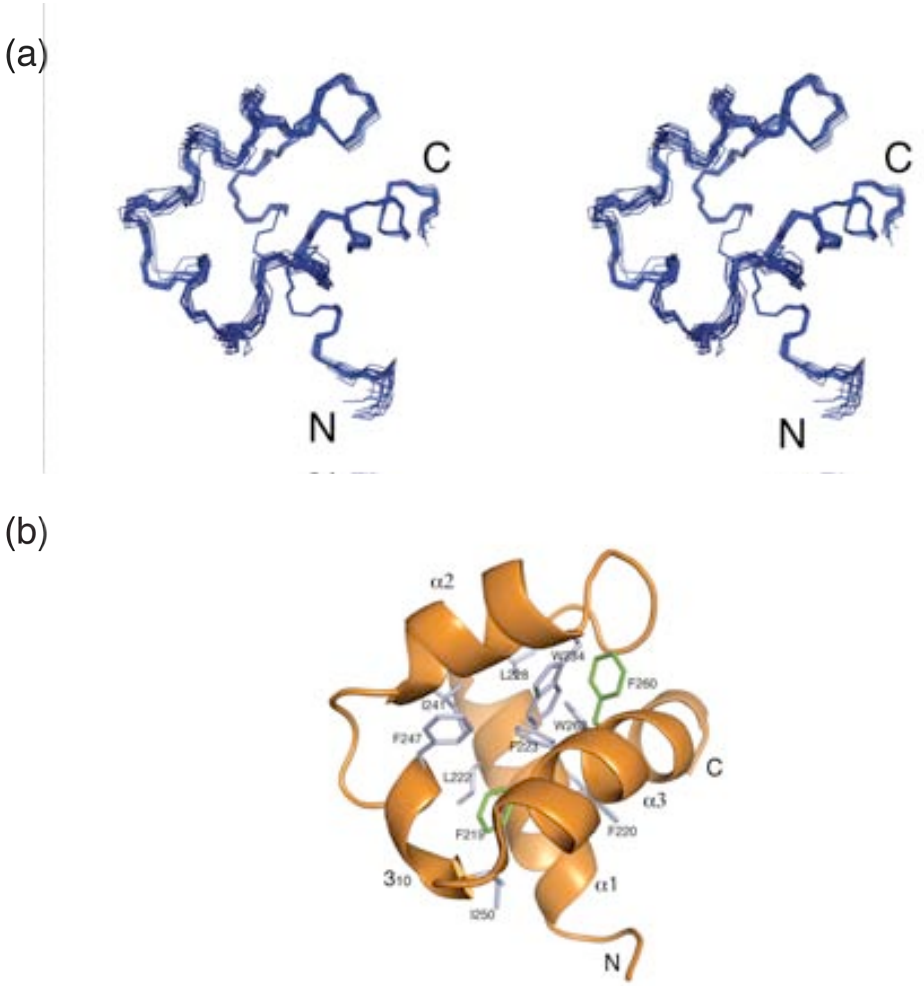



**Figure 17: URN1FF structure determined by NMR.**

**(a)** Stereo view of the overlay of the backbone atoms of URN1FF fifteen lowest energy structures after water refinement.

**(b)** Ribbon representation of the lowest-energy structure showing the side-chains for the hydrophobic residues that form the domain's core. The two conserved Phe that give name to the domain are depicted in green.

**Table 4: Structural statistics for the 15 lowest energy structures of URN1FF**

| **Restraints used for the calculation** <SA>**(a)** | |
|---|---|
| Intraresidual | 0 |
| Sequential ( \|i-j\| = 1) | 400 |
| Medium range ( 1< \|i-j\| ≤ 4) | 336 |
| Long range ( \|i-j\| > 4) | 561 |
| Unambiguous | all |
| Ambiguous | 0 |
| Dihedrals | 56 |
| Hydrogen bonds | 40 |
| All | 1297 |
| Restraint per residue ratio | 23.6 |
| **R.m.s. deviation (Å) from experimental (b)** | |
| NOE: | 0.02713 +/- $6.4 \times 10^{-4}$ |
| Bonds (Å) | 0.02252 +/-20.3 $\times 10^{-4}$ |
| Angles (º) | 1.684 +/- 0.1 |
| **Coordinate Precision (Å) (c)** | |
| backbone secondary structure elements | 0.16 |
| heavy atoms in secondary structure elements | 0.40 |
| heavy atoms all residues | 0.99 |
| **CNS potential energy (kcal mol$^{-1}$)** | |
| Total energy(d) | -42.86 +/- 4.5 |
| Electrostatic | -2286 +/- 58.7 |
| van der Waals | 255.6 +/- 27.7 |
| Bonds | 424.4 +/- 9.6 |
| Angles | 634.6 +/- 13.3 |
| **Structural quality (%residues)** 15 best structures | |
| in most favored region of Ramachandran plot | 86.7 |
| in additionally allowed region | 10.6 |

a) <SA> refers to the ensemble of the ten structures with the lowest energy

b) No distance restraint in any of the structures included in the ensemble was violated by more than 0.3Å

c) R.m.s. deviation between the ensemble of structures <SA> and the lowest energy structure

d) $E_{L\text{-}J}$ is the Lennard-Jones van der Waals energy calculated using the CHARMM-PARMALLH6 parameters. $E_{L\text{-}J}$ was not included in the target function during the structure calculation.

The overall structure of the URN1FF domain adopts the characteristic three-helical bundle fold, ($\alpha1$–$\alpha2$–$3_{10}$–$\alpha3$) as previously described for other FF domains (Allen et al, 2002; Gasch et al, 2006). The three α helices are formed by residues Glu215-Tyr226, Ser235-Glu242 and Asp253-Cys264 respectively, while the one turn $3_{10}$ helix spans the Asp246-Lys249 residues. The $\alpha1$ and $\alpha3$ helices are positioned close together in a nearly orthogonal manner, with several contacts between residues from each helix fixing this positioning (Glu215, Arg216, Phe219, Phe220 with Leu259, or Phe220, Phe223 and Asp224 with Trp263). As previously shown in the

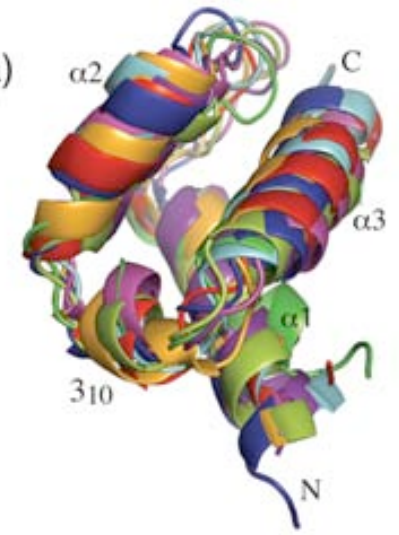
two other FF structures described, an extensive network of semi-conserved aromatic (Phe219, Phe220, Phe223, Trp234, Phe247, Phe260, Trp263) and aliphatic residues (Leu222, Leu228, Ile241, Ile250, Leu259) forms the core of the domain, with the two highly conserved Phe (Phe219 and Phe260) among them. The $3_{10}$ helix is anchored to the hydrophobic core by means of the contacts between Phe247 with residues in α1 (Leu 222) and α2 (Trp234, Ser238, Ile241) and those between Ile250 with residues in α1 (Glu215, Phe219) and α3 (Val255, Arg256, Leu259). The first loop (Lys226-Trp234) is the longest one in the structure. It is ordered due to the large number of long-range NOEs detected, such as Leu228 (a highly conserved aliphatic position in the FF family) with Trp263 and Lys230 with Trp263 and Cys264. Remarkably, Lys230 side-chain is packed to the domain and its Hζ protons display contacts to residues Leu228 and Gln237. Also, two hydrogen bonds within the loop (Leu228-Phe223, Asp229-Ser232) contribute to its rigidity. This rigidity of the loop is also present in Prp40FF1 and FBP11FF1 structures. Among the exposed residues, Asp244 and Asp246 are widely conserved. These residues are situated in the turn between α2 and $3_{10}$ helix and show NOEs to amino acids in the α1 helix (Ile218 and Leu222) suggesting a possible role in maintaining the overall fold, as described in the FBP11FF1 structure (Allen et al, 2002).

We superimposed the URN1FF domain structure determined in this work with all FF structures present to date in the Protein Data Bank (PDB) (Berman et al, 2000). Two sets of superimpositions were performed, a global fitting by default and a manual fitting restricted to the α1–α2–α3 secondary structure elements. All structural alignments were generated with CEalign (Shindyalov & Bourne, 1998). The superimposition of the distinct FFs to URN1FF is good in all cases (Fig. 18 (a)), with RMSD values below 2 Å when secondary structural elements are superimposed, despite the low sequence identity among the structures, (Fig. 18 (b)). In all cases the region comprising the first loop and the α2 -$3_{10}$ helices is the most variable, while α1 and α3 superimposed best.

*Analysis of the charge distribution on the surface of FF domains*

FF domains have distinct overall $pK_a$ values (Gasch et al, 2006; Sun et al, 2004). Prompted by this observation we set to investigate the electrostatic charge distribution of the known FF structures to determine if $pK_a$ values could correlate with the conservation of charged patches on the surface. Electrostatic surface representations of FF structures are shown in Fig. 19, ordered according to their overall $pK_a$ values, basic (A), neutral (B) and acid (C). FBP11FF1 and CA150FF1 ($pK_a$=9.9, and 9.2 respectively) have an extended and coincident basic area centred around α3. CA150FF2 ($pK_a$=9.1) has its positively charged patch fragmented by small negative ones.

(a)




(b)

| | RMSD/Å | | |
|---|---|---|---|
| | SS (optAlign) | Global (Cealign) | seq identity % |
| scPrp40FF1 | 1.88 | 2.21 | 14.5 |
| hsFBP11FF1 | 1.32 | 1.52 | 16.4 |
| hsFBP11FF5 | 1.34 | 1.95 | 12.7 |
| hsCA150FF1 | 1.48 | 1.94 | 20.4 |
| hsCA150FF2 | 2.15 | 2.61 | 21.8 |
| hsCA150FF3 | 1.73 | 2.26 | 30.9 |
| hsCA150FF4 | 1.74 | 2.43 | 16.4 |

**Figure 18: Comparison of URN1FF to other determined FF structures.**

**(a)** URN1FF domain (in *green*) superimposed to the lowest energy structure of the different FF domains for which the structure is deposited in the PDB. Color codes for the remaining FF domains are scPrp40FF1-red, hFBP11FF1-violet, hFBP11FF5- pale green, hCA150FF1-light blue, hCA150FF2-dark blue, hCA150FF3-pink, hCA150FF4-orange. The corresponding PDB codes are scPrp40FF1-2B7E, hFBP11FF1-1UZC, hFBP11FF5-2CQN, hCA150FF1-2DOD, hCA150FF2-2E71 hCA150FF3-2DOE, and hCA150FF4-2DOF.

**(b)** R.M.S.D. of the different FF domains with respect to the URN1FF. Fittings were generated taking either the secondary structure elements (α1–α2–α3 helices) or the whole molecules (Global).

CA150FF4 ($pK_a$=6.1) and FBP11FF5 ($pK_a$=6.8) surfaces have positive and negative electrostatic patches distributed on the surface but located at different positions. CA150FF3 ($pK_a$=6.9) has a basic patch centred at α2. This patch is partially present in CA150FF4 but absent in FBP11FF5 although they all have similar $pK_a$ values. Furthermore, URN1FF and Prp40FF1 have similar acidic $pK_a$ values (4.5 and 4.7 respectively), but their surfaces are different. Both domains contain an extended negatively charged but it is not located in the same area (in URN1FF it involves the region including α1 and α3 helices while in Prp40FF1 the patch also comprises the loop between $3_{10}$ and α3). Hence, FF domains have noticeable charged patches on the surface but their localization is variable even for domains with similar $pK_a$ values.

To test the hypothesis that having similar $pK_a$ values but different charge distribution may have implications in ligand binding, we titrated the URN1FF domain with the TPR motif shown to bind to Prp40FF1 (Gasch et al, 2006). No changes on the URN1FF amides were appreciated upon addition of up to five times excess of the TPR motif (Fig. 20). Apparently, non-conservative changes in the domain interacting surface (Thr163, Arg164 and Lys179 in

Prp40FF1 correspond to Glu, Asn and Glu in the URN1FF) preclude the interaction. Since the TPR binding site contains several glutamic acid residues (Gasch et al, 2006), the lack of binding to URN1FF domain may imply that electrostatic repulsion could have prevented binding.

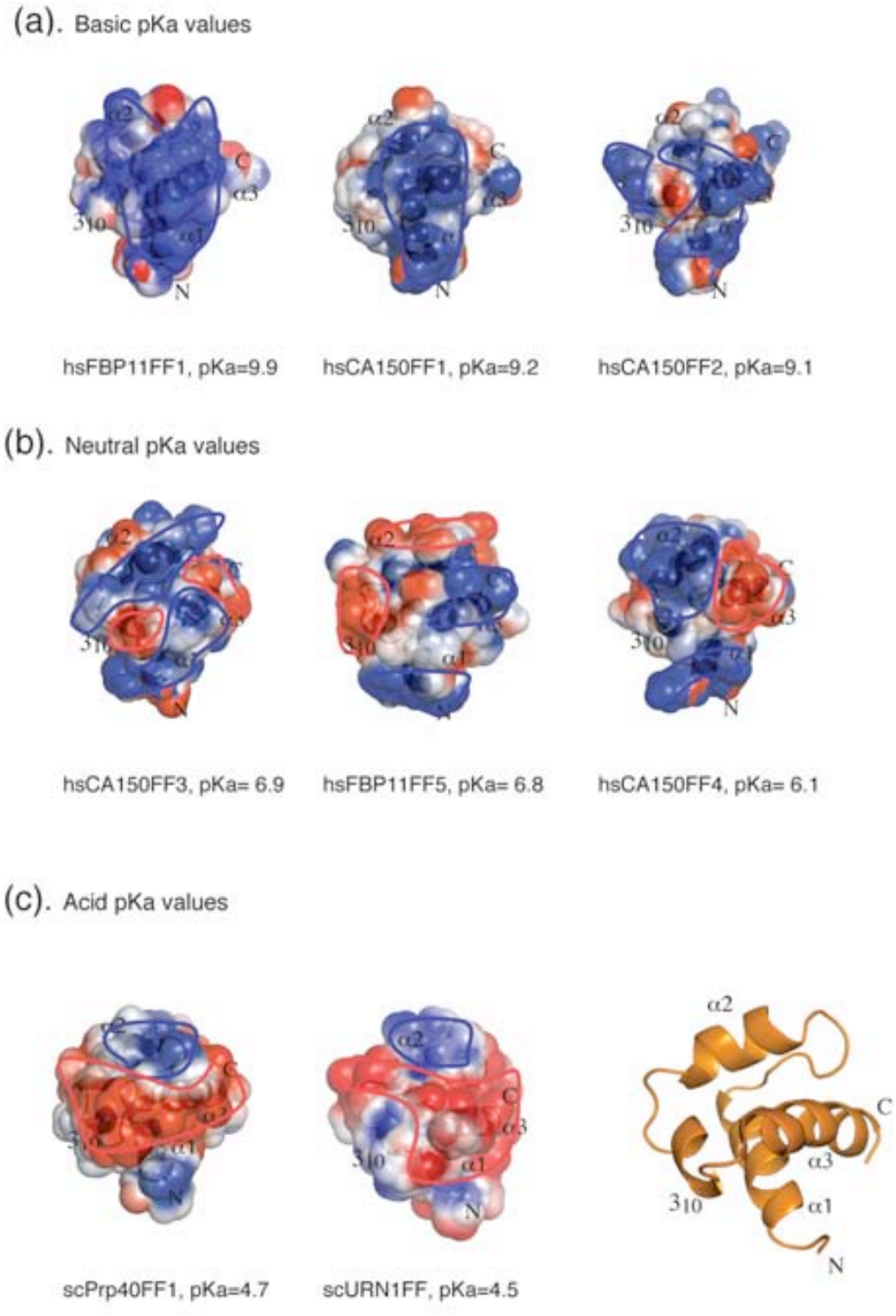



**Figure 19: Charge distribution for the different FF domains.** All electrostatic surface representations were calculated with APBS. They are shown as semitransparencies to allow recognition of the secondary

structural elements. To facilitate their identification protein names and pKa values are labelled on top of each surface. The electrostatic patches are marked with either a blue (positive) or red (negative) contour. (A) FF domains with basic pKa values: FBP11FF1, CA150FF1 and FF2. (B) FF domains with neutral pKa values: FBP11FF5, CA150FF3 andFF4. (C) FF domains with acid pKa values: Prp40FF1 and URN1FF. To further facilitate the identification of secondary structural elements, a cartoon representation is shown next to the surface of the URN1FF domain.

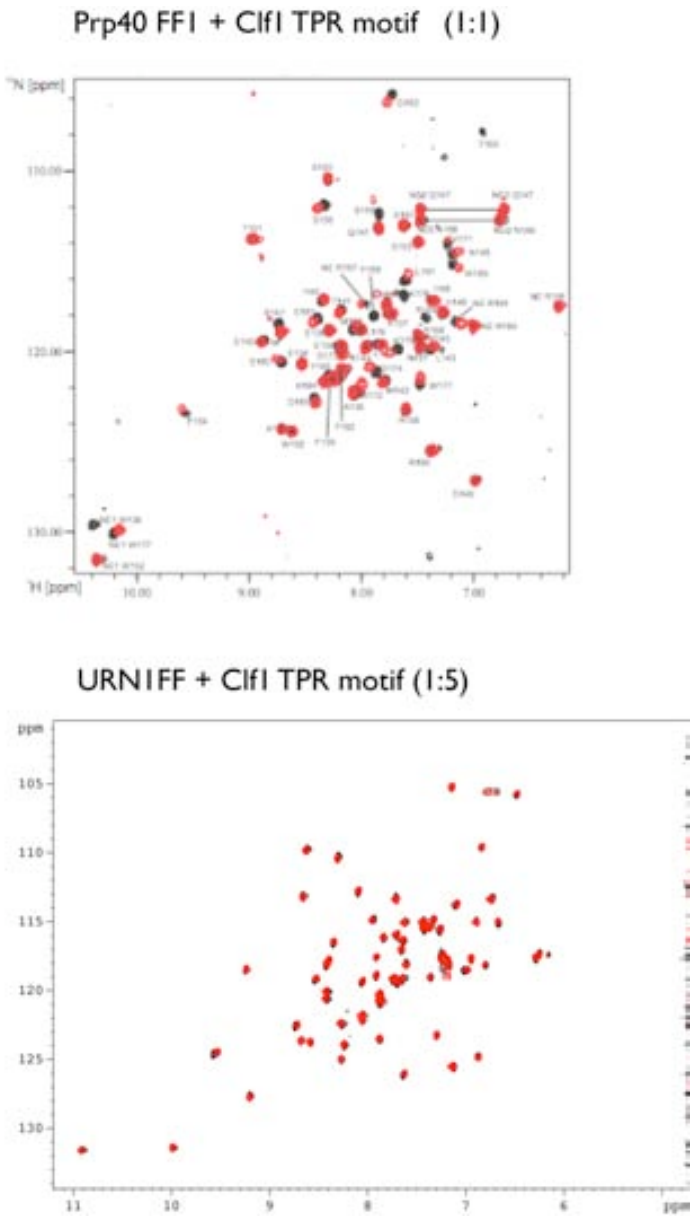



**Figure 20: Binding of Prp40FF1 and URN1FF to Clf1 TPR motif.**

*Up*- $^{15}$N-HSQC spectra of $^{15}$N labelled Prp40FF1 domain in the absence (black) and presence (red) of unlabelled Clf1-TPR motif. *Down*- $^{15}$N-HSQC spectra of $^{15}$N labelled URN1FF domain in the absence (black) and presence (red) of unlabelled Clf1-TPR motif.

*Comparison to other proteins with similar fold*

In the Structural Classification Of Proteins (SCOP) database (Murzin et al, 1995) FF domains are described as having a 3-helix bundle fold and are grouped together with the protein super-families of the C-terminal domain of Ser/Thr phosphatase 2C, the B-form DNA mimic viral Ocr protein, the helical domain present in the chromatin associated DEK protein and the IscX-like proteins, which are predicted to function as iron donors in the assembly of iron-sulfur clusters. By looking beyond the structures in SCOP, we realized that FF domains also share a similar architecture with SURP domains. SURP domains, however, are classified in the SCOP database as the only members of the SURP module fold, described as an irregular array of 5 helices. Nevertheless, as is explained in the original work (Kuwasako et al, 2006), only four helices correspond to the domain, and the fifth one corresponds to a N-terminal extension. SURP domains can therefore be reconsidered as having the $\alpha1-\alpha2-3_{10}-\alpha3$ topology.

*a-. Comparison at the sequence level*

Since the proteins grouped with FFs in SCOP are functionally unrelated and have distinct global architecture with respect to FF domains, we did not investigate conservation at the sequence level. In contrast, SURP domains have been shown to interact with other splicing factors, although there are some indications that involve SURP domains in RNA interactions (Kuwasako et al, 2006). Thus, the similar architecture with respect to FF domains and the appearance of SURP domains in splicing associated proteins (as some FF domains, including URN1FF) suggest a potential structural similarity between both domains. To explore the possibility of detectable sequence similarity, we performed BLASTp (Altschul et al, 1990) searches, using either SURP or FF domains as baits to ascertain whether they might be capable of retrieving either FF or SURP domains, respectively. This approach has been previously used to detect other divergent domains (Castresana & Saraste, 1995; Pascual et al, 1997). However, in our case, such searches only selected members of same family.

The high degree of sequence divergence is also reflected when we attempt to compare FF and SURP sequences, whose coordinates are deposited in the PDB, using ClustalX (Thompson, 1997). The independently aligned FF and SURP sequences were re-aligned with the profile-alignment ClustalX option, yielding a comparison that does not maintain the integrity of the secondary structure elements (data not shown). Thus, we carried out the sequence comparison manually by aligning the $\alpha1$ and $\alpha3$ helices of URN1FF and SF3a120SURP1 on the basis of their structural superimposition. As displayed in Fig. 21, both sets of sequences do not have conserved residues in common. Even critical residues involved in the definition of the protein core or in the definition of the secondary structure are not conserved

between both sets of domains. The fact that they share a related fold but no sequence conservation may imply that different contacts to ligands are possible in order to perform a similar functional role (such as protein or RNA recognition).

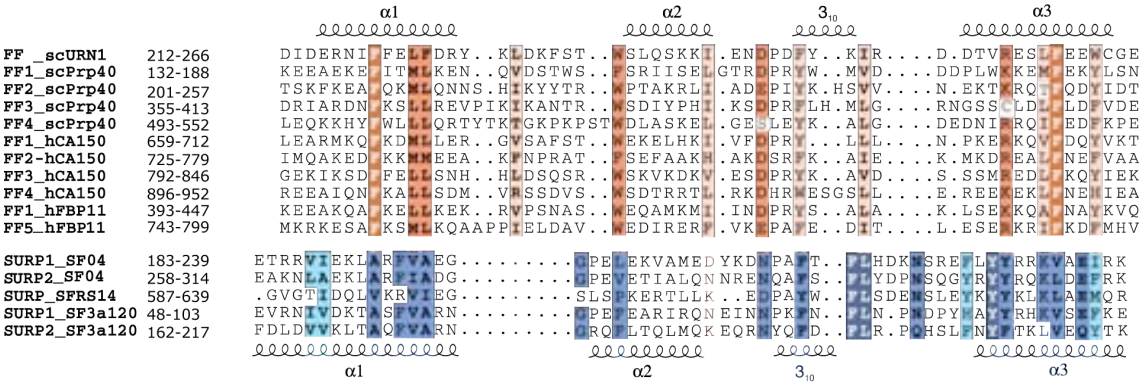



**Figure 21: Structure-based alignment of FF and SURP aligned profiles.**

Conserved residues between the two profiles are indicated with *red gradient* (FFs) and *blue gradient* (SURPs) backgrounds according to the degree of conservation (dark is the most conserved). For instance, in the FF family the Phe in the first helix (the one corresponding to the first Phe in the name FF) is conserved in all sequences, while the SURP family contains an Ala at this equivalent position. Both families have a pair of hydrophobic semi-conserved residues in position +3 with respect to this Phe/Ala highly conserved residue. Similarly in the last helix there are four positions conserved in each family, but not between the families. Secondary structure elements for each profile and residue numbers of the distinct domains are indicated.

*b-. Comparison at the structural level*

With the idea of detecting possible binding site conservation at the structural level, we superimposed the URN1FF domain to the SCOP hits and to SURP domains. The structural alignments were generated either with a global fitting of the secondary structure elements or with local fittings concentrated in the first and third helices, which show a higher degree of conservation among FF domains (Gasch et al, 2006).

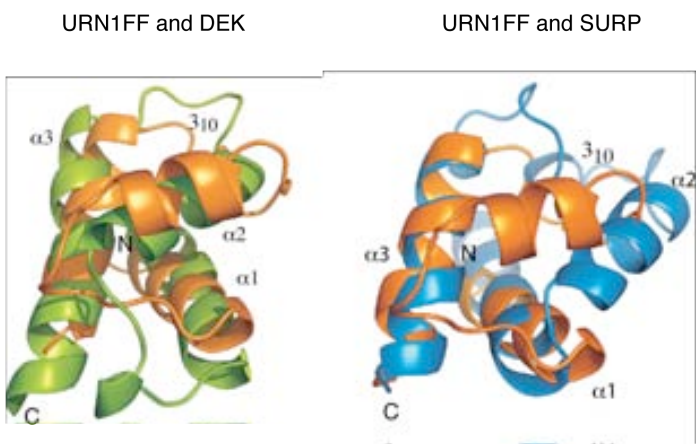
The family of the C-terminal domain of Ser/Thr phosphatase 2C protein has already been compared with FF domains when the structure of the FBP11FF1 domain was reported, and the authors concluded that the structural similarity could be a reflection of functional relatedness since the FBP11FF1 domain was shown to interact with a phosphorylated ligand (Allen et al, 2002). The URN1FF structure compares with the C-terminal domain of the phosphatase with a RMSD value between both domains of 4.45 Å. With the DEK C-terminal domain and with the

IscX protein, the global fitting is also of the same range. Selecting only α1 and α3 improves the fitting to 2.10 Å, 2.41 Å and 3.15 Å respectively (Fig. 22 (a) *left*).

The Ocr viral protein has a remarkably different architecture from that of the FF domain (α1–$3_{10}$–α2–α3–$3_{10}$–α4–α5). We tried to optimize the regions for the comparisons but we could only get RMSD values close to 3.5 Å when we performed a global alignment, fitting elements of secondary structure to residues present in loops. Thus this match seems to be an artefact of the comparison more than a genuine hit indicative of structural similarity.

In the comparison with SURP domains, we obtained an RMSD value of 4.5 Å for the global fitting. The RMSD decreased to a value of 1.8 Å (Fig. 22 (a) *right*) when only α1 and α3 helices were used. From this superimposition it becomes clear that the overall fold of the two families is similar. However, the loop connecting the α1 and α2 helices is much longer in FF domains while in SURP domains, the longest loop connects the $3_{10}$ to α3 (Fig. 22 (a) *right*). These differences probably explain the variable orientation of α2 and the $3_{10}$ helices between the two families and account for the large RMSD value observed in the global fitting. All RMSD values are collected in Fig. 22 (b).

(a).




(b).

| | RMSD/Å | | | | |
|---|---|---|---|---|---|
| | C-term, S/T phosphatase 2C | B-form DNA-mimic Ocr | DEK C-term | IscX protein | SF3a120SURP1 |
| **SS (MolMol)** | 2.10 | 8.80 | 2.41 | 3.15 | 1.63 |
| **SS (optAlign)** | 2.63 | 8.07 | 2.60 | 3.60 | 1.80 |
| **Global (cealign)** | 4.45 | 3.60 | 3.79 | 3.86 | 4.48 |

**Figure 22: Structural comparison between URN1FF and other related structures.**

**(a)** *Left and Right*: URN1FF domain (in *orange*) superimposed to either DEK C-terminal structure (in *green, PDB code* 1Q1V) or the SF3a120SURP1 structure (in *blue, PDB code* 2DT6). Representations show the superimpositions according to the α1–α3 fittings using the optAlign alignment script. Secondary structure elements are indicated.

**(b)** RMSD of the family representatives for the *Another 3-helical bundle fold*, as well as for the

SF3a120SURP1 domain (in *bold*) with respect to the URN1FF. Fittings were generated taking either the secondary structure elements α1–α3 helices or the whole molecules (Global). *PDB codes*: C-terminal S/T phosphatase 2C-1A6Q, B-form DNA-mimic Ocr-1S7Z, DEK-C-terminal-1Q1V, IscX protein-1UJ8, SF3a120SURP1-2DT6.

Overall we show that both α1 and α3 orientations are well conserved in the FF domains, the SCOP targets and the SURP domains investigated. The main differences are concentrated in the α2 and the $3_{10}$ helices as well as in the length of the first and last loops.

### 4.1.4 DISCUSSION

FF domains are highly divergent in sequence and this is reflected in the variety of ligands that they can bind to. Since many FF domains are charged, $pK_a$ values have been used to interpret some of the interactions with phosphorylated charged ligands. The comparisons here presented suggest that this parameter should be used with caution for this purpose. Conserved patches are related to sequence conservation and not to the presence of a given number of charged residues (responsible for the theoretical calculation of $pK_a$ values). Besides, the presence (or absence) of charged patches could explain why Prp40FF1 domain can bind to a TPR repeat while URN1FF domain is unable to do so albeit both domains having an acidic $pK_a$ value. It also explains why FBP11FF1 and probably some CA150 FF domains can bind to certain phosphorylated ligands while the Prp40FF1 domain cannot (Gasch et al, 2006). The presence of charged patches has been observed in inositol binding domains such as the spectrin PH domain or in nucleotide binding domains such as DEP domains. In these cases, additional contacts from semi-conserved aromatic residues were also identified (Civera et al, 2005; Hyvonen et al, 1995; Macias et al, 1994).

To complicate matters further, the binding data obtained for one given domain cannot be directly extrapolated to other FF sequences (apart from the same FF domain of a different specie) or from yeast to humans, since the yeast proteins do not have direct orthologues. The phylogenetic reconstructions performed by Sun *et al.* with CA150 and FBP11 proteins and by Gasch *et al.* using only FF domains showed this difficulty (Gasch et al, 2006; Sun et al, 2004). Thus, if the FF domain sequences are not evolutionary related, knowing a given $pK_a$ value may not have a direct correlation with the presence (or absence) of charged patches and more importantly with the binding capabilities of the particular FF domain.

By comparing structural and biological information from proteins with similar characteristics, we have come across a parallelism between FF and SURP domain families. From a structural point of view, both are all-helical domains, have a α1–α2–$3_{10}$–α3 topology and, as shown by our structural analysis, give similar RMSD values when superimposed to each

other as those of FF domains to the rest of the other members detected in the SCOP database. From this perspective, we suggest that SURP modules should be included in the FF fold.

We observed that the main difference between all these structures resides in the orientation of the second helix. The sequence variability observed for α2 may be the consequence of ligand specialization or on the contrary, the result of limited evolutionary pressure only aimed towards maintaining a helical secondary structure. The first explanation may account for the variable ligands that can be recognized by the FF fold where the second helix plays a role in binding, (dsDNA recognition in the DEK domain, protein binding in SF3a120SURP2 and Prp40FF1 domains). The second option will explain the binding described for FBP11FF1 where only residues located in α1 and α3 are involved in phospho-peptide recognition. A representation of these binding sites is shown in Fig. 23.

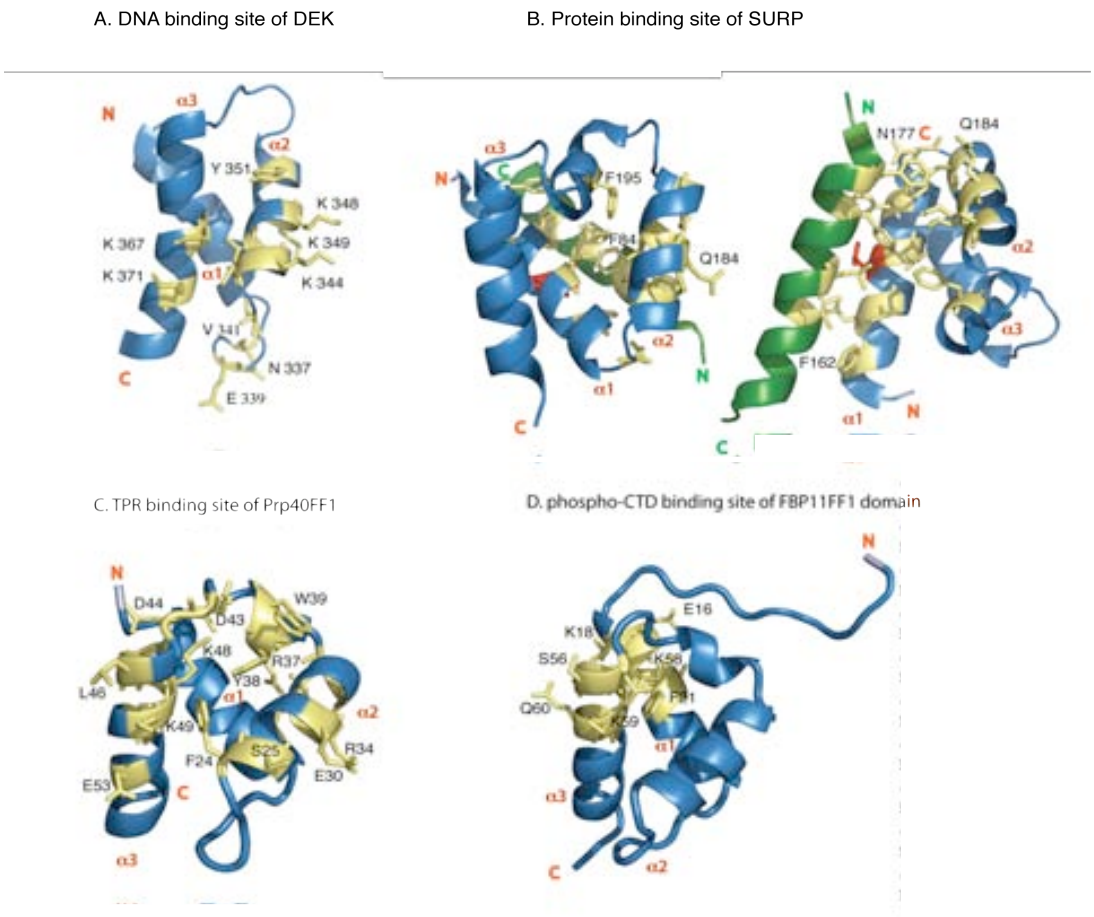



**Figure 23: Binding sites and complexes for representative sequences with the FF fold.**
Binding sites described for the DEK-C -terminal, SF3a120SURP, Prp40FF1 and FBP11FF1 domains are displayed in panels A, B, C and D respectively. All domains are represented using the same orientation of the molecule, to facilitate the visual structural comparison of binding sites. Secondary structural elements are displayed as cartoons and labelled in red. Since there is only the structure of the SF3a120SURP complex, the remaining representations show the domain residues affected upon binding. In all cases

these residues are labelled and displayed as light yellow coloured sticks. In the DEK and FBP11FF1 domains most of the affected residues are charged residues. For the SURP complex structure, both domain and ligand residues affected upon binding are shown. The Leucine critical for the interaction is shown in red. To facilitate the inspection of the interacting site we also include a rotated view of the complex. In this case the interacting residues are mainly hydrophobic. The Prp40FF1 binding site involves both charged and hydrophobic residues.

Certainly additional experiments aimed to structurally characterize complexes where FF domains are involved may help to clarify how FF domains can recognize ligands and their role in the proteins where they are present.

## 4.2 SECTION 2

### 4.2.1 INTRODUCTION

Prp40 is a splicing factor essential for yeast viability originally identified as a suppressor of 5' end U1 RNA point mutations (Kao & Siliciano, 1996). Prp40 is a U1 snRNP- associated protein that participates in the early steps of yeast pre-mRNA splicing. Prp40 associates with the branch-binding point (BBP) protein to bring the 5' splicing site and the intron branch point into spatial proximity (Abovich & Rosbach, 1997). This function is mediated by the interaction of the WW domains of Prp40 with a PPxY motif present in the BBP (Branch Binding Protein). WW domains also mediate the binding of Prp40 to Prp8, a component of U5 snRNP (Abovich & Rosbach, 1997). Additionally, Prp40 binds the phosphorylated C-terminal domain (phospho-CTD) of RNA polymerase II through regions involving the WW and FF domains (Morris & Greenleaf, 2000). However, a subsequent study on the structure of the Prp40 WW domain pair also showed that in the absence of additional FF domains the pair of WW domains does not interact with a doubly phosphorylated-CTD repeat (Wiesner et al, 2002).

The solution structure of the first FF domain of Prp40 has also been determined (Gasch et al, 2006). That study also examined the binding of Prp40FF1 to the splicing factor Clf1 and to a doubly phosphorylated tandem CTD repeat. The Prp40FF1 folds as an all-helical domain with a $\alpha1$-$\alpha2$-$3_{10}$-$\alpha3$ architecture. The binding site for the association with the first TPR motif of Clf1 involves the helices $\alpha2$, the $3_{10}$ and the N-terminal half of $\alpha3$. In contrast, no interaction was detected for the Prp40FF1 domain with the phospho-CTD repeats and for the Prp40FF4 domain with the TPR motif of Clf1 (Gasch et al, 2006).

Recently, other ligand partners have been identified for Prp40 FF domains, namely Snu71, a component of U1 snRNP, and also Luc7, a splicing factor associated with U1 snRNP and involved in the 5' splicing site recognition (Ester & Uetz, 2008; Fortes et al, 1999; Gottschalk et al, 1998). Furthermore, it was shown that only the region comprising the first two FF domains of Prp40 is critical for yeast viability, while deletion of the region including FF3 and FF4 results in a slow growth phenotype (Ester & Uetz, 2008).

Of the FF2, FF3 and FF4 domains, we were able to obtain a stable sample only for the last one. Indeed, several constructs spanning FF2 and FF3 and even that of the FF1-2 pair were either partially structured or unstable after refolding from inclusion bodies.

Thus, we focused the structural work on the Prp40FF4 domain. Furthermore, prompted by the observation that the charge distribution of the FBP11FF1 region involved in the interaction with the phospho-CTD repeats is partially conserved in Prp40FF4, we also examined whether this domain also interacted with the phospho-CTD repeats, but no binding was detected under our experimental conditions.

### 4.2.2 EXPERIMENTAL PROCEDURES

*Sample preparation*- The Prp40FF constructs corresponded to residues 134-260 (FF1-2), 200-260 (FF2), 346-412 (FF3) and 488-552 (FF4). The fragments were amplified by PCR using genomic DNA of *Saccharomyces cerevisiae* as template and cloned into a pETM30 vector for the production of a N-terminal His-tag followed by GST and a TEV protease cleavage site. For the Prp40FF4 domain production, *Escherichia coli* BL21 (DE3) cells were grown at 37ºC and induced with 0.4 mM IPTG for 3 hours. The domain was purified as described previously (Bonet et al, 2008). $^{15}$N- and $^{13}$C- labelled protein was prepared following the method developed by Marley and co-workers (Marley et al, 2001) using D-[$^{13}$C] glucose and $^{15}NH_4Cl$ as sole sources of carbon and nitrogen respectively.
The Prp40FF4 sample for NMR experiments was concentrated to 0.5mM in 20mM sodium phosphate buffer, 130mM NaCl, 0.03% (w/v) $NaN_3$ in 90% $H_2O$, 10% $D_2O$ or 100% $D_2O$ at pH 5.2.

*NMR spectroscopy*- NMR spectra were recorded at 285K on either a Bruker DRX-600 or Bruker DRX-800 NMR spectrometer. Resonance assignments were obtained combining the information from standard triple resonance experiments (CBCA(CO)NH, HNCBCA, HCCCONH, HCCH-TOCSY) and homonuclear 2D TOCSY (65 ms mixing time) and 2D NOESY (120 ms mixing time) experiments. Intra- molecular proton distance restraints were obtained from peaks assigned in 2D-NOESY, $^{15}$N-NOESY and $^{13}$C-NOESY experiments. All spectra were processed with the NMRPipe/NMRDraw software (Delaglio et al, 1995) and analyzed with CARA (Bartels et al, 1995).
For the NMR titrations with the phospho-CTD, $^{15}$N-Prp40FF4 was prepared at 0.2mM and unlabelled ligand was added to a final ratio 1:3.

*Structure calculation*- Distance restraints derived from fully assigned peaks in NOESY experiments, $^3J(H^N, H^\alpha)$ coupling constants and hydrogen bond restraints were used for structure calculation. The structures were calculated with the program CNS (Brünger et al, 1998) and ARIA 2.0 (Habeck et al, 2004), using 7 iterations of 20 structures and a final iteration of 80 structures. Water refinement was applied to the 30 lowest energy structures of the final iteration. The 15 lowest energy structures were analyzed with PROCHECK-NMR (Laskowski et al, 1996) and the statistics from the analysis are shown in Table 5. MOLMOL (Koradi et al, 1996) and PyMOL (DeLano, 2002) were used to visualize the structures and generate the figures.

*Peptide synthesis*- The sequence SY**pS**PT**pS**PSY**pS**PT**pS**PSY corresponding to a doubly phosphorylated pair of repeats of the C-terminal domain of RNA polymerase II was manually

synthesized using Fmoc solid phase peptide synthesis and Rink amide matrix (Wellings & Atherton, 1997). The peptide was cleaved from the resin with a TFA 95% / $H_2O$ 2.5% / TIS 2.5% mixture, precipitated in cold ether and purified by HPLC in a C4 reverse phase column (Vydac).

*Accession numbers*- The chemical shifts have been submitted to the BioMagResBank (BMRB accession number 16176) and the protein coordinates to the Protein Data Bank (PDB ID code 2kfd)

### 4.2.3 RESULTS AND DISCUSSION

We determined the solution structure of the yeast Prp40 FF4 domain using standard multidimensional heteronuclear NMR spectroscopy. The final ensemble of 15 lowest energy conformers is displayed in Fig. 24 and the structural statistics are summarized in Table 5.

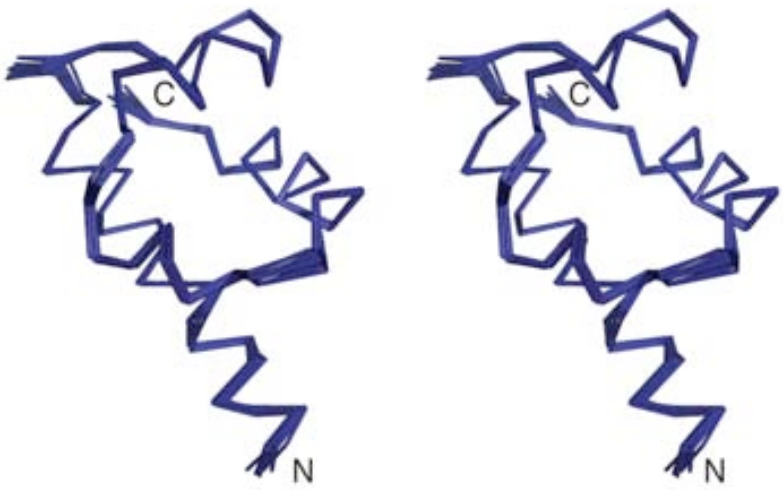



**Figure 24: Stereo view of Prp40FF4 structures.**

Overlay of the backbone 15 lowest energy conformers of the Prp40FF4 domain after water refinement.

Like other previously described FF domains (Allen et al, 2002; Bonet et al, 2008; Gasch et al, 2006), Prp40FF4 folds as a compact four-helical bundle, with a α1-α2-$3_{10}$-α3 topology (Fig. 25 (a)). The helices are formed by residues Glu489-Thr507 (α1), Trp519-Leu526 (α2), Tyr532-Gly536 ($3_{10}$) and Asp539-Phe549 (α3). Each helix is connected to the rest of the helices through long-range NOEs. For instance, Tyr532, located in the $3_{10}$ helix, is in contact with residues in α1 (Phe500 and Leu503) α2 (Ser523) and α3 (Arg542). Most of the residues

involved in these contacts are well conserved among FF domains and form the hydrophobic core of the protein (Fig. 25 (b)). In addition, $D_2O$ exchange experiments showed that amide resonances in the region comprising the C-terminal part of helix α1 (Trp501-Tyr508) and helix α3 (Ile541-Phe549) have the slowest rate of exchange with the solvent while α2 and $3_{10}$ are the most accessible.

**Table 5: Structural statistics for the 15 lowest energy structures of Prp40FF4.**

| | |
|---|---|
| **Restraints used for the calculation** <SA>**(a)** | |
| Intraresidual | 0 |
| Sequential ( \|i-j\| = 1) | 846 |
| Medium range ( 1< \|i-j\| ≤ 4) | 782 |
| Long range ( \|i-j\| > 4) | 936 |
| Unambiguous | All |
| Ambiguous | 0 |
| Dihedrals | 82 |
| Hydrogen bonds | 62 |
| All | 2564 |
| Restraint per residue ratio | 37.1 |
| **R.m.s. deviation (Å) from experimental (b)** | |
| NOE: | $0.01216 \pm 3.4 \times 10^{-4}$ |
| Bonds (Å) | $0.01309 \pm 5.1 \times 10^{-4}$ |
| Angles (º) | 21.21 ± 1.11 |
| **Coordinate Precision (Å) (c)** | |
| backbone secondary structure elements | 0.26 |
| heavy atoms in secondary structure elements | 0.87 |
| heavy atoms all residues | 1.06 |
| **CNS potential energy (kcal $mol^{-1}$)** | |
| Total energy(d) | -2327.6 ± 413.6 |
| Electrostatic | -2862 ± 64.2 |
| van der Waals | -393.9 ± 35.7 |
| Bonds | 39.2 ± 4.4 |
| Angles | 184.1 ± 6.2 |
| **Structural quality (%residues)** 15 best structures | |
| in most favored region of Ramachandran plot | 86.9 |
| in additionally allowed region | 10.8 |

a) <SA> refers to the ensemble of the ten structures with the lowest energy

b) No distance restraint in any of the structures included in the ensemble was violated by more than 0.3Å

c) R.m.s. deviation between the ensemble of structures <SA> and the lowest energy structure

d) $E_{L-J}$ is the Lennard-Jones van der Waals energy calculated using the CHARMM-PARMALLH6 parameters. $E_{L-J}$ was not included in the target function during the structure calculation.

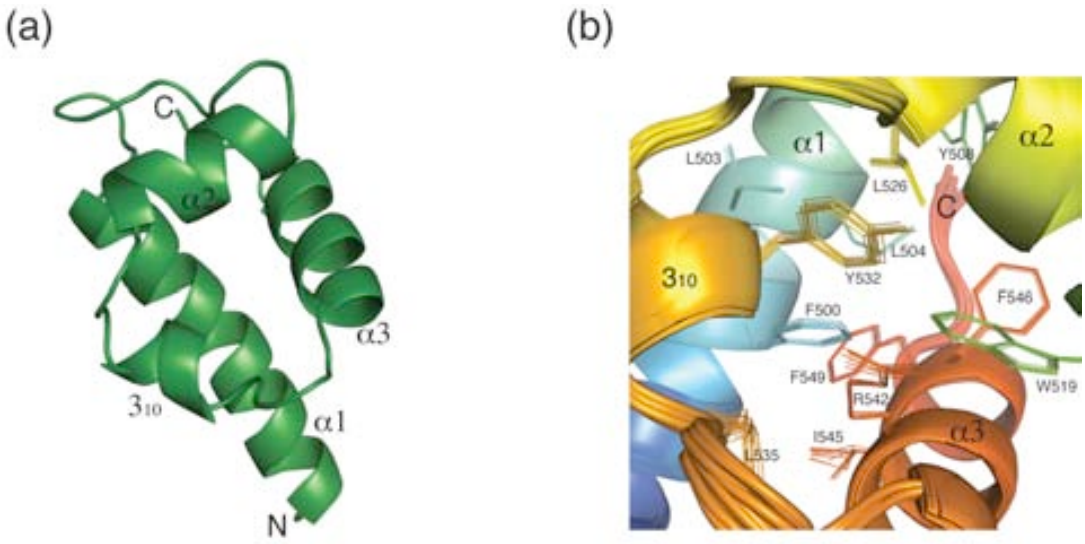



**Figure 25: Solution structure of the Prp40FF4 domain.**

**(a)** Ribbon representation of the lowest-energy structure of Prp40FF4.

**(b)** Detailed view of the hydrophobic core of Prp40FF4. The residues involved in the formation of the core are as follows: Phe500, Leu503 and Leu504 from α1, Thr507 and Tyr508 from the first loop, Trp519 and Leu526 from α2, Tyr532 and Leu535 from $3_{10}$ and Arg542, Ile545, Phe546 and Phe 549 from α3.

The loop connecting the first two alpha helices exhibits distinct features in Prp40FF4 compared to other FF structures. A superimposition of Prp40FF4 to Prp40FF1 and URN1FF shows that the loop in Prp40FF4 presents an extra turn, as a consequence of the insertion of several additional amino acids in the FF4 sequence (superimposition of structures and sequences shown in Fig. 26 (a) and Fig. 26 (b) respectively). Although the loop is longer than in other structures its flexibility is not increased. It presents a high number of detectable NOE peaks with residues in the alpha helices, i.e., Lys515 side-chain with the rings of Trp501 (in α1) and Phe546 (in α3) or Pro516 with Ala522 (in α2) and Phe546. Remarkably, the hydroxyl groups of Thr507 and Tyr508 (in the beginning of the loop) participate in a network of NOEs with methyl groups of Leu503, Leu504 and side-chains of Pro516, Glu525 and Leu526 respectively. Accordingly, both side-chains are well defined in the structure.

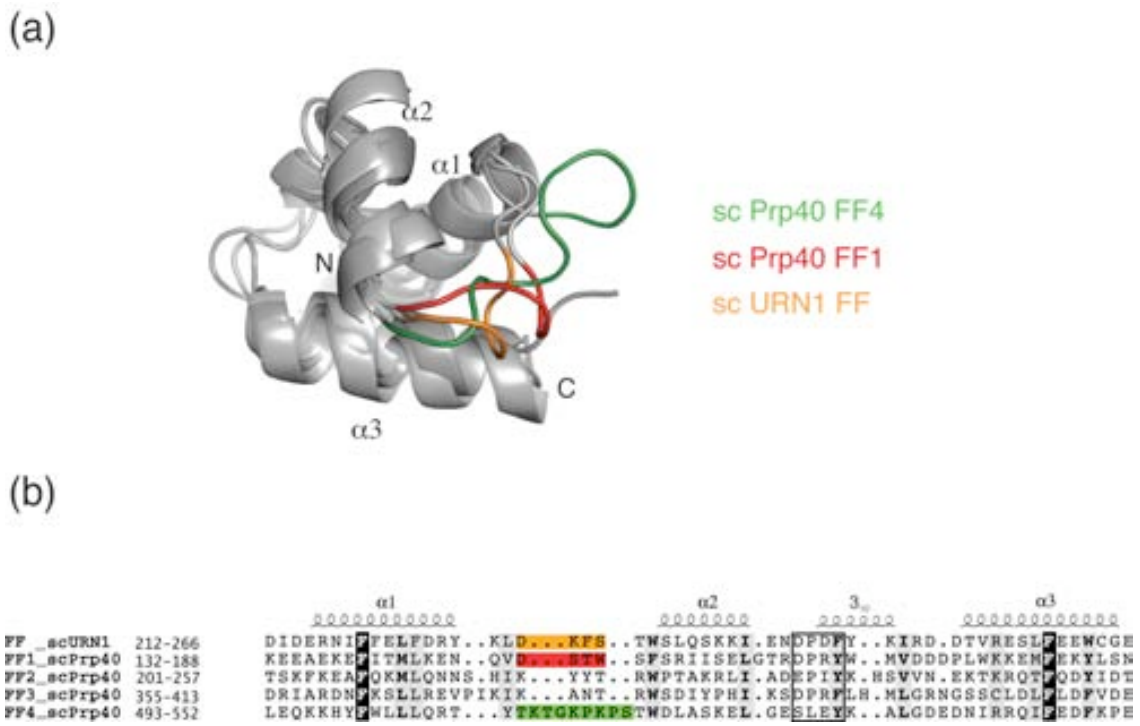



**Figure 26: Comparison of Prp40FF4 domain with other yeast FF domains.**

**(a)** Superimposition of Prp40FF4 to Prp40FF1 and URN1FF structures. The region corresponding to the loop connecting α1 and α2 is coloured in green for Prp40FF4, in red for Prp40FF1 and in orange for URN1FF.

**(b)** Sequence alignment of *Saccharomyces cerevisiae* FF domains. The alignment was generated with ClustalX (Thompson, 1997) and edited manually. Conserved residues are shaded in grey. The region corresponding to the first loop is shaded in colours according to the superimposed FF domains of Fig. 26 (a). The DPR(Y/F) motif is boxed.

Notably, the motif DPR(Y/F), often present in FF domains in the loop connecting α2 and $3_{10}$, correspond to SLEY in the FF4 domain sequence. This motif is involved in the binding of Prp40FF1 to a TPR repeat of Clf1 and its absence in Prp40FF4 was related to the inability of the domain to interact with the same TPR repeat (Gasch et al, 2006).

Like Prp40FF1, Prp40FF4 showed no binding to a tandem repeat of the phospho-CTD in our experimental conditions (Fig. 28). The phospho-CTD binding site in FBP11FF1 was located in a positively charged region comprising the N-terminal of helices α1 and α3 (Allen et al, 2002). In Prp40FF1, this region is mainly negatively charged and therefore unfavourable for the phospho-CTD binding. Nevertheless, in Prp40FF4 the region comprising the N-terminal of helices α1 and α3 is rather positively charged, so in this case the lack of binding to the phospho-CTD cannot be explained based on the surface electrostatic potential (Fig. 27).

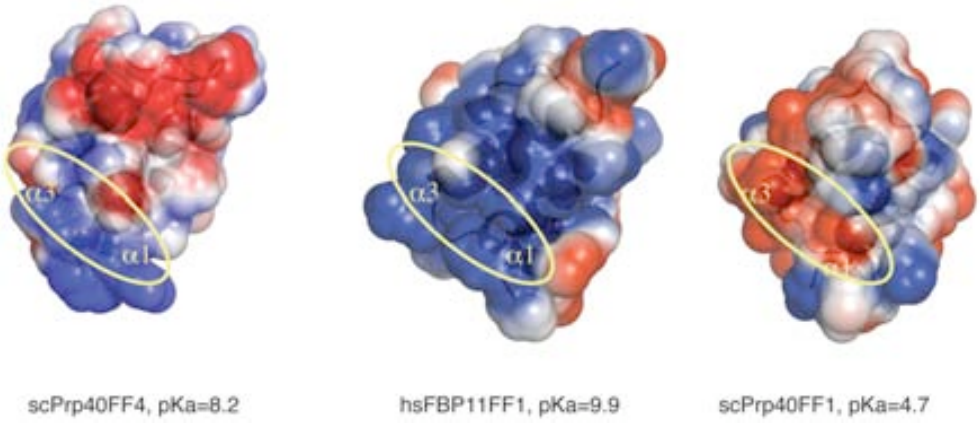



**Figure 27: Electrostatic surface plots of Prp40FF4, FBP11FF1 and Prp40FF1, calculated withAPBS (Baker et al, 2001).** Positive and negative surfaces are drawn in blue and red respectively. The region corresponding to the phospho-CTD binding surface on the FBP11FF1 domain is marked with a yellow circle on the three domains. pKa values for each domain are indicated.

Unfortunately, we were unable to obtain stable samples of the FF2 and FF3 constructs because of their insolubility and high degree of precipitation after refolding procedures. We obtained only a soluble sample for the construct including FF1 and FF2 domains but it turned out to be only partially folded (Fig. 28). Still, we titrated it with increasing amounts of the tandem repeat of phospho-CTD, but we did not detect chemical shift changes in the amide resonances of the FF1-2 domain. Although we cannot conclude that the Prp40FF2 domain does not interact with the phospho-CTD due to the intrinsic low dispersion of the amide resonances, it seems that the addition of the phospho-CTD does not favour the folding and stabilization of the Prp40FF1-2 construct. Possibly, other regions of Prp40 such as the remaining FF2 and/or FF3 domains, or the presence of additional factors may be required to facilitate the interaction of Prp40 with the phospho-CTD repeats of RNA polymerase II.

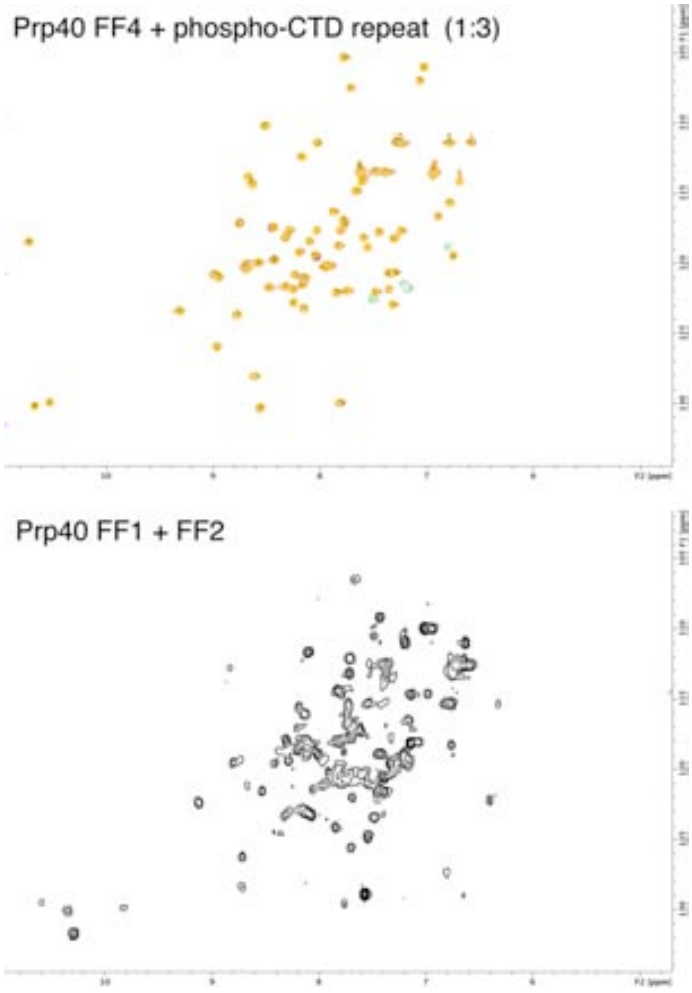



**Figure 28: Interaction of Prp40 FF domains with the phospho-CTD.**

*Up*- $^{15}$N-HSQC spectrum of Prp40FF4 domain after the addition of up to 3-fold excess of a double repeat of the phospho-CTD.

*Down*- $^{15}$N-HSQC spectrum of the partially folded Prp40 FF1+2 construct.

# CHAPTER 2

## BINDING STUDIES OF CA150 FF DOMAINS TO THE PHOSPHORYLATED C-TERMINAL DOMAIN (CTD) OF RNA-POLYMERASE II

## 5.1 INTRODUCTION

The first specific function assigned to a FF domain was the binding to the C-terminal domain of the RNA polymerase II (CTD), a repeated motif of the sequence YSPTSPS that serves as a platform for the recruitment of nuclear factors. Carty et al. reported that the FF domains of the transcription elongation repressor CA150 interacted with the hyperphosphorylated (on serine 2 and serine 5 residues of the consensus sequence) form of the CTD but not with an unphosphorylated form (Carty et al, 2000). This interaction is necessary for CA150-mediated transcription repression, since a truncated CA150 lacking all six FF repeats exhibited a reduced capacity to repress $\alpha_4$-integrin compared to the wild-type (Goldstrohm et al, 2001). Besides, the binding data for the CA150 FF domains indicated that only FF2 and FF5 interacted with the phospho-CTD, suggesting that the CA150 FF domains display different binding specificities (Carty et al, 2000). After, though, it was found that *in vivo*, the region including the FF6 domain was essential for binding to transcription/splicing factors, including the phospho-CTD. The study also unveiled the importance of FF domains for co-localization of CA150 to nuclear speckles (Sanchez-Alvarez et al, 2006).

Pull-down experiments were used to identify multiple new interacting partners for the CA150 FF domains. A wide peptide-array based analysis of the CA150FF1-3 association to one of these ligands, Tat-Sf1, resulted in the identification of the sequence $(D/E)_{2/5}$-F/W/Y-$(D/E)_{2/5}$ as a consensus binding motif for the FF domains (Smith et al, 2004). A binding mechanism was proposed, in which FF domains associate with ligand targets through multiple weak interactions. Data also indicated that distinct FF domains possessed similar affinities, that each domain had only one recognition site and that it did not appear to be cooperative effects. Furthermore, binding assays of the CA150 FF1-3 construct to a synthetic phospho-CTD peptide confirmed that the doubly phosphorylated CTD is a preferred target for the FF domains (Smith et al, 2004).

Regions containing WW and FF domains from yeast Prp40 splicing factor were also reported to associate with the phospho-CTD. In this case, however, it seemed that the presence of both domains simultaneously was required for the interaction (Morris & Greenleaf, 2000). This observation was further confirmed by the absence of binding detected to the phospho-CTD for the pair of WW and the individual FF1 and FF4 domains from Prp40 (Gasch et al, 2006; Wiesner et al, 2002).

The first FF domain from HYPA/FBP11, that represented the first determined structure for an FF domain, also showed ability to bind to a tandem repeat of the phospho-CTD. Moreover, the interacting region for the phospho-CTD within the FF surface was mapped to a cluster of positively charged amino acids located at the N-terminal of helices $\alpha$1 and $\alpha$3 (Allen et al, 2002).

In contrast, the structural determinants within the CA150 FF domains for the phospho-CTD binding have not been established yet. In addition, from the different published studies it is not clear whether all the CA150 FF domains are capable of interacting with the phospho-CTD or if they display different affinities. Hence, we decided to produce each of the six FF domains of CA150 to examine the binding to a synthetic tandem repeat of the phospho-CTD and to determine the regions within the FF domain implicated in the interaction. Further, we tested the interaction of a construct including the first two FF domains to compare the binding to that of the individual domains.

Of the six individual FF domains, FF5 appeared to be only partially folded and FF6 was largely unfolded. Conversely, the other fourth FF are structured domains. Our results indicate that individual CA150 FF1-5 domains bind to the phospho-CTD, with a rather weak affinity, as judged by NMR titrations. Additionally, these HSQC experiments also indicate that the binding affinity of a given FF is not improved by the presence of other FF domains. Furthermore, the mapping of the residues that experienced changes upon phospho-CTD addition reveals that the region involved in the binding is not exactly the same for the distinct CA150 FF domains.

## 5.2 EXPERIMENTAL PROCEDURES

*Sample preparation*- The CA150 FF constructs corresponded to residues 650-715 (FF1), 720-783 (FF2), 650-783 (FF1-2), 785-853 (FF3), 890-956 (FF4), 951-1009 (FF5), 1010-1077 (FF6) and 951-1077 (FF5-6). Fragments were amplified by PCR from human cDNA clone CA150-pEFBOS and sub-cloned into either pETM10 and pETM11 vectors for the generation of a protein with a N-terminal His-tag or a N-terminal His-tag followed by a TEV protease cleavage site, respectively. For the over-expression of the FF domains, *Escherichia coli* BL21 (DE3) cells were grown at 37ºC or 20ºC and induced with 0.4 mM IPTG. Domains were purified, generally, using Ni Sepharose (GE Healthcare). After cleavage with TEV protease domains were further purified using gel-filtration chromatography on Superdex™ 75 prepgrade (GE Healthcare). $^{15}$N- and $^{13}$C- labelled protein was prepared following the method developed by Marley and co-workers (Marley et al, 2001) using D-[$^{13}$C] glucose and $^{15}NH_4Cl$ as sole sources of carbon and nitrogen respectively. Sample for NMR experiments typically contained 0.5mM of FF domain in 20mM sodium phosphate buffer, 130mM NaCl, 0.03% (w/v) $NaN_3$ in 90% $H_2O$, 10% $D_2O$ or 100% $D_2O$ at pH 6.5.

*NMR spectroscopy*- Backbone experiments (CBCA(CO)NH and HNCBCA) for the assignment of the amide resonances from the distinct FF domains were recorded either at 285K or 298K on

a Bruker Advance III-600 spectrometer. Spectra were processed with the NMRPipe/NMRDraw software (Delaglio et al, 1995) and analyzed with CARA (Bartels et al, 1995).
NMR titrations with the phospho-CTD were monitored by HSQC experiments, using $^{15}$N-labelled CA150 FF domains prepared at around 0.2 mM and adding unlabelled ligand to a final ratio 1:5.

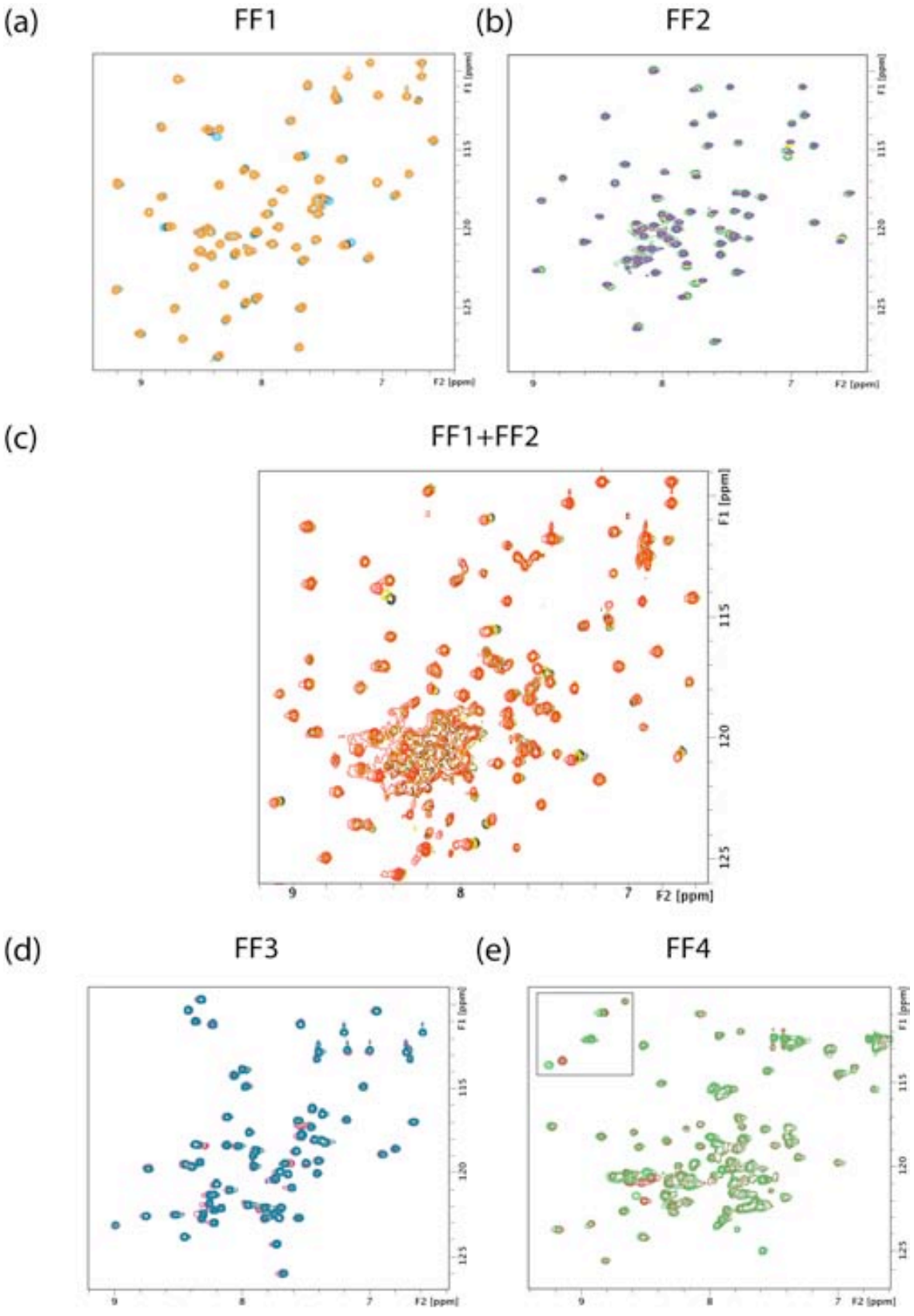
*Peptide synthesis*- The sequence SY**pS**PT**pS**PSY**pS**PT**pS**PSY corresponding to a doubly phosphorylated pair of repeats of the C-terminal domain of RNA polymerase II was manually synthesized using Fmoc solid phase peptide synthesis and Rink amide matrix (Wellings & Atherton, 1997). The peptide was cleaved from the resin with a TFA 95% / $H_2O$ 2.5% / TIS 2.5% mixture, precipitated in cold ether and purified by HPLC in a C4 reverse phase column (Vydac).

## 5.3 RESULTS AND DISCUSSION

*Interaction studies of the CA150 FF domains with the phospho-CTD*

We produced large quantities of $^{15}$N- and $^{15}$N/$^{13}$C- samples of the individual CA150 FF1-4 domains. NMR measurements indicated that the fourth FF domains were folded (Fig 29 (a-e)). In contrast, even though we used several distinct strategies of expression and purification, we could not obtain stable samples for the FF5 and FF6 domains, which largely aggregate during purification and subsequent concentration of the samples. Furthermore, $^{15}$N-HSQC experiments showed a limited number and a poor dispersion of amide resonances, indicating that the two domains were basically unstructured proteins (Fig. 29 (f,g)). We also attempted to purify a construct including both domains, but it behaved in a way similar to the individual domains. In fact, the solution structures of the CA150 FF1 to FF4 domains, but not those of the FF5 and FF6, have been deposited in the Protein Data Bank (PDB codes: FF1-2dod, FF2-2e71, FF3-2doe, FF4-2dof) by the Riken Structural Genomic/Proteomics Initiative (RSGI). Nevertheless, as the NMR assignments for the domains are not available, we recorded backbone experiments (CBCACONH and HNCBCA) for the CA150 FF1 to FF4 domains in order to map their amide resonances. We could not measure backbone experiments for the FF5 and FF6 domains due to their insufficient concentration and little stability.
NMR titrations were done by adding increasing amounts of a stock solution of a phospho-CTD double repeat (peptide sequence SY**pS**PT**pS**PSY**pS**PT**pS**PSY) to $^{15}$N-labelled samples of the distinct CA150 FF domains at a final ratios 1:3 or 1:5 (protein:ligand). We observed that in all cases the amide resonances of the domains experienced changes upon addition of the phospho-

CTD, except for the FF6 (Fig.29), where the poor signal dispersion detected in the HSQC spectrum precluded a detailed analysis.



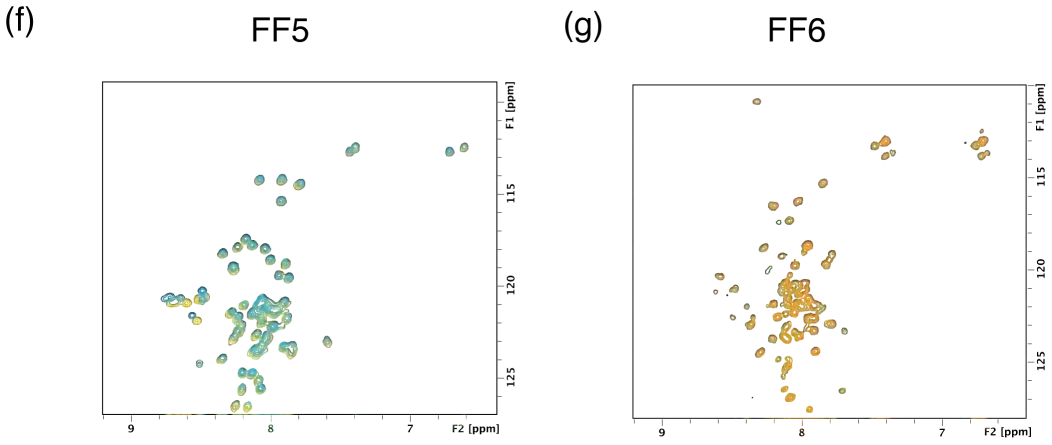



**Figure 29: NMR titrations of CA150 FF domains with the phospho-CTD.**

**(a-g)** Overlay of $^{15}$N-labelled CA150 FF domains 1-6 in absence and presence of unlabelled phospho-CTD double repeats. Protein:ligand molar ratios corresponded to 1:3 for FF1, FF2, FF1-2, FF3 and FF6 titrations and to 1:5 for FF4 and FF5 titrations. Spectra were recorded at 298K for FF1-4 domains and at 285K for FF5 and FF6 domains. For the FF4 domain, the region corresponding to the tryptophan indol protons is shown zoomed in a box at the upper left corner.

Our experiments indicate that the interaction of all CA150 FF domains with the phospho-CTD is similar. In all cases, the binding of the CA150 FF domains to the phospho-CTD appears to be weak, as judged by the small number of residues affected and the small changes in the chemical shifts of their amide resonances. In the case of FF5 and FF6 the presence of the phospho-CTD did not result in stabilization or even a partial folding of the domains (as is sometimes the case for disordered proteins upon binding to ligands (reviewed in (Dyson & Wright, 2002)), indicating that the phospho-CTD does not induce significant conformational changes in the tertiary structure of these two domains.

The weak interaction that we detected for all the CA150 FF domains with the phospho-CTD is in concordance with the binding mode observed for the CA150 FF domains to ligands (Smith et al, 2004). In that study it was reported a low and roughly equal affinity for the interaction of individual CA150 FF domains 1-3 to their binding motifs, with values of $K_d$ generally greater than 100 μm. However, it was discussed that the binding to the phospho-CTD repeats might be stronger, based on the value of $K_d \sim 50$ μm obtained for the FBP11FF1 domain interaction with a double repeat of the phospho-CTD (Allen et al, 2002). Our results, though, point to an interaction between the CA150 FF domains and the phospho-CTD that is rather weak, more in accordance with the binding mode observed for CA150 FF1-3 to the distinct motifs. We did isothermal titration calorimetry (ITC) in order to determine the $K_d$ for the CA150 FF2 domain interaction with the phospho-CTD but, unfortunately, we could not determine a confident value due the low affinity of the binding, thereby suggesting that CA150

FF domain interactions with the phospho-CTD are weaker compared to that of the FBP11FF1 domain. Furthermore, backbone assignment allowed us to map the regions within the CA150 FF domains 1 to 4 involved in the phospho-CTD (Fig.30). We could not map the amide resonances of FF5 and FF6 because of the impossibility to obtain good backbone spectra.

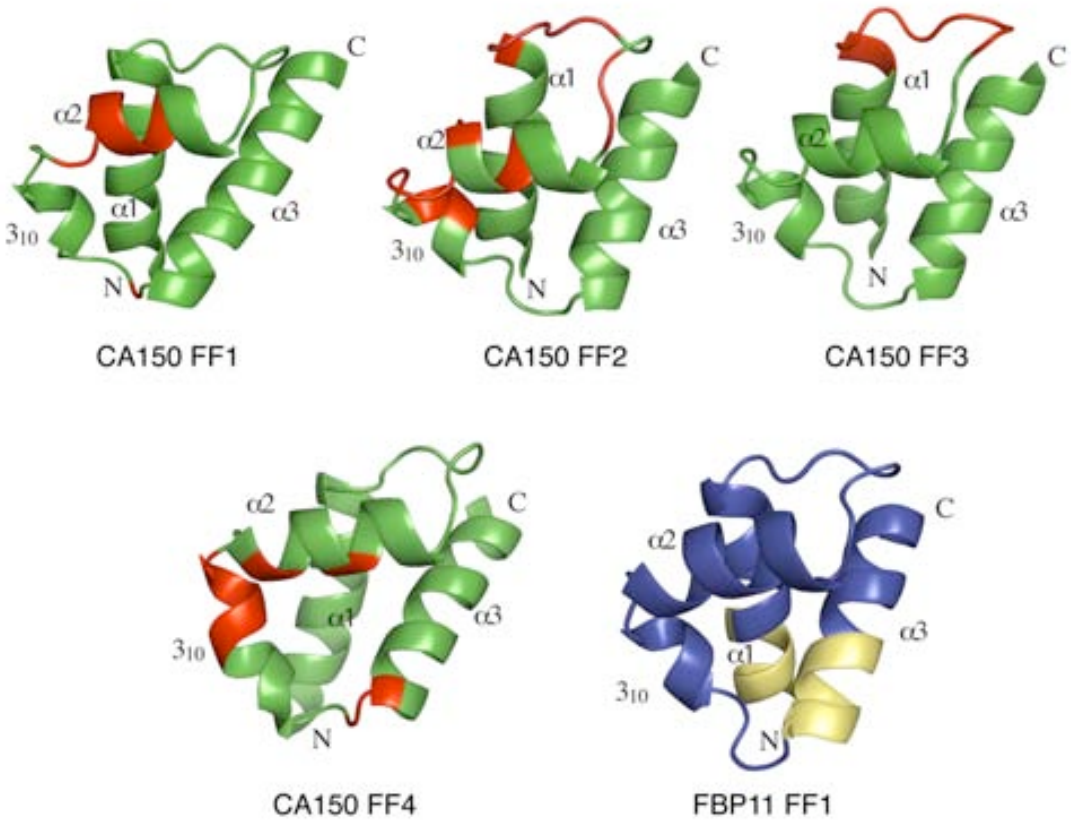



**Figure 30: phospho-CTD interacting regions within the CA150 FF1-4 (in green) and FBP11FF1 (in blue) domains.** Binding sites are indicated in red for CA150 FF domains and in yellow for the FBP11 FF1. Secondary structure elements are labelled.

We found that the interacting surfaces are mainly located at the end of α1, the first loop, α2 and $3_{10}$ helices, but are not equal for the different domains. In particular, in FF1 and FF3 few residues participate in the interaction, and are located at helix α2 and the first loop, respectively. In FF2 and FF4, the interaction involves a higher number of residues and comprises a more extended area, including parts of the first loop, α2 and $3_{10}$ helices. Remarkably, the phospho-CTD interacting surfaces found for the CA150 FF1-4 domains greatly differ from that reported for the FBP11FF1, which was mapped at the N-terminal of α1 and α3 helices. In the case of the FBP11FF1, the interaction involved a cluster of positively charged residues within the FF

surface including two conserved lysine residues. Conversely, the binding sites in the CA150 FF do not map to regions that contain large basic patches. Rather, they contain solvent-exposed aromatic/aliphatic residues that form accessible hydrophobic patches (Fig. 19). Consequently, it seems that additional contacts other than electrostatic interactions are involved in the CA150 FF interaction with the phospho-CTD. These observations also confirm that the overall $pK_a$ value for a given FF domain is not a good indicative of its binding preferences, because for the CA150 FF domains, their overall $pK_a$ values do not correlate with the electrostatic nature of their interacting surfaces.

Our results illustrate that, apart from the wide range of specificities that seems to exist for the FF domains, they can also display different binding sites for the interaction with the same ligand, for instance, the phospho-CTD. Notably, the CA150 FF region for the phospho-CTD binding include, generally, the first loop, the $\alpha 2$ and the $3_{10}$ helices, which on the basis of domain superimpositions are the most variable parts among FF domains (Fig. 18) and also when comparing FF domains to other members of the FF fold (according to SCOP database classification) (Fig. 22). Furthermore, in the interactions in which the FF fold members participate, the $\alpha 2$ helix plays a role (Fig. 23). Thus, the observed binding region of CA150 FF domains to the phospho-CTD would coincide better with other binding sites described for members of the FF fold than that of the FBP11FF1.

Another aspect that we examined was whether the interaction of the FF domains with the phospho-CTD was enhanced in the presence of several domains. Hence, we performed the titration with a construct including the FF1 and the FF2 domains of CA150. We observed that the changes upon binding in the double construct map approximately in the same region that for the individual domains, and that the apparent affinity of binding is similar, even slightly smaller for the tandem construct (Fig. 31). Our results, which indicate that the presence of two FF domains does not improve the binding to the phospho-CTD, are thus in agreement with the study of Smith et al. that reported that CA150 FF domains did not exhibit any cooperative effect in ligand binding (Smith et al, 2004).

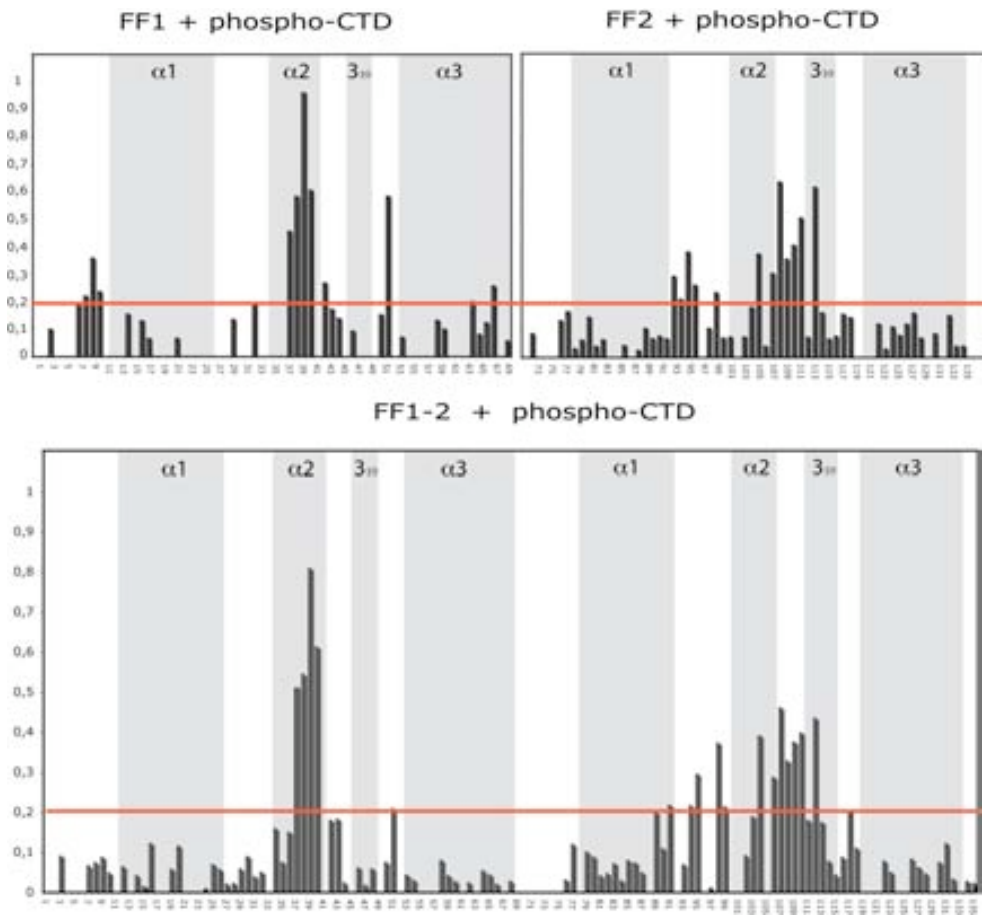



**Figure 31: Interaction of individual and tandem FF1 and FF2 domains with the phospho-CTD.** Bar representation of the chemical shift changes in the amide resonances of the individual CA150 FF1 and FF2 domains and the double construct FF1-2. Chemical shifts were calculated as the difference between the value at 1:3 protein:ligand ratio and the value for the free domain, using the equation $\Delta\delta^{av} = ((\Delta\delta_{1H})^2 + (\Delta\delta_{15N}/5)^2)^{1/2}$. Significant changes are considered for residues whose average chemical shift difference is more than 0.2 ppms between the free and bound state.

In conclusion our study has revealed that all CA150 FF1 to 5 domains possess ability to interact with a double repeat of the phospho-CTD. FF6 might also bind the phospho-CTD, although we have not detected the interaction surely due to the few and averaged shifts observed for the amides of the domain. Moreover, as judged by the $^{15}$N-HSQC experiments, the affinity of the protein:peptide interaction is low (and roughly equivalent for all the domains), and it does not increase in the presence of several domains, thus reinforcing the view that FF domains bind to ligands in a non-cooperative form. Finally, we observed that the interacting region within the CA150 FF domains is distinct from the previously reported binding site of the FBP11FF1.

# CHAPTER 3

NMR STRUCTURE OF THE p190-A RhoGAP FF1 DOMAIN AND INSIGHTS INTO ITS PHOSPHORYLATION MECHANISM

## 6.1 INTRODUCTION

RhoGTPases are a family of proteins with key roles in the regulation of actin cytoskeleton dynamics (reviewed in (Heasman & Ridley, 2008)). These proteins are found in an active, GTP-bound form, or in an inactive GDP-bound form. The switch between these two states is controlled by three protein families, one of which comprises the RhoGTPase activating proteins (GAPs).

RhoGAPs are classified into a large number of subfamilies on the basis of sequence homology studies and domain composition. Among these, the members of the p190 RhoGAP subfamily comprise the only cytoplasmatic proteins that contain FF domains. This subfamily is formed by just two proteins, p190-A and p190-B, which have a common domain distribution, with an N-terminal GTPase domain followed by four FF domains in the central part and a C-terminal RhoGAP domain. p190 RhoGAPs participate in a variety of processes such as angiogenesis, cell migration, neuronal morphogenesis and cancer cell invasion (Mammoto et al, 2009; Tcherkezian & Lamarche-Vane, 2007)).

p190-A RhoGAP was first identified as a protein associated with p120 RasGAP that is rapidly phosphorylated on Tyr in fibroblasts stimulated with epidermal growth factor (Ellis et al, 1990). In fact, phosphorylation is a prevalent mechanism for the regulation of p190-A RhoGAP activity. For instance, the interaction of p190-A RhoGAP with p120 RasGAP is strongly modulated by c-Src phosphorylation on Tyr1105 (Roof et al, 1998). Tyr1105 is also the preferred target for the action of Abl-related gene (Arg) and Breast tumor (Brk) kinases to regulate p190-A RhoGAP mediated processes such as neuronal morphogenesis in the postnatal brain or tumor malignancy in breast cancer cells, respectively (Hernandez et al, 2004; Shen et al, 2008).

Phosphorylation on another Tyr residue of p190-A RhoGAP also appears to be crucial in regulating its interaction with TFII-I (Jiang et al, 2005), a signal-induced transcription factor implicated in distinct processes such as the transcriptional regulation of the c-fos gene and the G-kinase signal transduction pathway (reviewed in (Roy, 2007)). That study reported that p190-A RhoGAP interacts via its FF domains with the N-terminal of TFII-I. This region of TFII-I, which includes a putative leucine zipper, is responsible for the binding to Butron's tyrosine kinase (Btk) (Sacristan et al, 2004) and is required for dimerization of TFII-I (Cheriyath & Roy, 2001). The binding to p190-A RhoGAP prevents TFII-I translocation into the nucleus and subsequent transcriptional activity. Furthermore, the phosporylation of platelet-derived growth factor (PDGF) receptor $\alpha$ kinase on Tyr308 (located at the first FF domain of p190-A RhoGAP) is sufficient to disrupt the interaction and restore the capacity of TFII-I to enter the nucleus (Jiang et al, 2005). This is the only reported case of phosphorylation on an FF domain as a way of modifying its ligand binding specificity. However, the structural

determinants within the FF domain implicated in the interaction with TFII-I and the potential molecular rearrangements caused by the phosphorylation are unknown.

To better understand some of these aspects, we solved the structure of the first FF domain of p190-A RhoGAP (herein referred to as RhoGAPFF1) using multidimensional heteronuclear Nuclear Magnetic Resonance (NMR). The structure of RhoGAPFF1 presents several differences compared to other previously characterized FF structures. Among these, RhoGAPFF1 has a α1-α2-α3-α4 architecture instead of the α1-α2-$3_{10}$-α3 classical arrangement.

Moreover, the structure revealed that the phosphorylation target Tyr308 participates in the formation of the hydrophobic core of the domain, and consequently, it is an unfavourable position to be phosphorylated. We thus examined whether Tyr308 could be efficiently phosphorylated. The phosphorylation of RhoGAPFF1 required a partial unfolding of the domain, which occurred at 37˚C. In addition, a phospho-mimicking Y308D mutant was mostly insoluble under similar expression conditions as the wild type, thereby suggesting that the presence of a charged residue at this position is incompatible with the FF fold.

Finally, under our experimental conditions, p190-A RhoGAPFF1 interacted only very weakly with the N-terminal of TFII-I, even at high excess of the latter. Similarly, the fourth FF domain of p190-A RhoGAP, which appeared to be unfolded on the basis of our NMR experiments, also displayed very weak binding to TFII-I. However, we cannot exclude that the feasibility of the interaction depends on additional events, such as phosphorylation on TFII-I. This possibility would resemble the interaction of FF domains with the CTD repeat, which occurs only when the repeat is phosphorylated in Ser residues (Carty et al, 2000; Morris & Greenleaf, 2000).

## 6.2 EXPERIMENTAL PROCEDURES

*Sample Preparation*- The constructs of the FF1 (267-331 and 267-327) and FF4 (485-540) domains of p190-A RhoGAP and the fragments of the N-term TFII-I (1-89, 1-60, 21-89 and 21-60) were obtained by PCR amplification from human cDNA clones KIAA1722 (Kazusa Human cDNA Project) and IRATp970H0799D (imaGenes) and sub-cloned into a modified pET24-d vector that generates a N-terminal fusion His-tag followed by the thioredoxin protein and a TEV protease cleavage site.

The RhoGAPFF1Y308D mutant was obtained by PCR-mutagenesis using the *wt*RhoGAPFF1 domain as the template.

Production and purification of all the constructs was performed as described previously (Bonet et al, 2008). However, the FF1 domain of RhoGAP contain three His residues in C-terminal positions that allowed it to bind the $Ni^{2+}$ resin in absence of the His-tag. Thus, after cleavage with TEV, the fusion protein was directly separated from the domain with a Superdex$^{TM}$ 30 column (GE healthcare Life Sciences). $^{15}N$ and $^{13}C$ -labelled samples were produced using the method developed by Marley and coworkers (Marley et al, 2001). All samples were prepared in 50mM sodium phosphate buffer, 150mM NaCl, 0.5mM $NaN_3$ in 90% $H_2O$, 10% $D_2O$ or 100% $D_2O$ at pH 7.2.

*NMR data assignment*-All NMR experiments for the structure determination were recorded at 285K in a Bruker Advance III-600, processed with the NMRPipe/NMRDraw software (Delaglio et al, 1995) and analyzed with XEASY (Bartels et al, 1995). Triple resonance experiments CBCA(CO)NH, HNCBCA, H(CCCO)NH, HCCH-TOCSY and 2D-TOCSY were used for $^{1}H$,$^{15}N$ and $^{13}C$ backbone and side- chain assignments. 2D-NOESY, $^{15}N$-NOESY and $^{13}C$-NOESY experiments were used to obtain the intra-molecular proton distance restraints. All NOESY experiments were acquired with mixing times of 120 ms. Spectra used for the calculation were integrated with the batch integration method of the XEASY package.

*NMR titration experiments*-For the $^{15}N$-HSQC experiments, $^{15}N$-labelled samples were prepared at 0.3 mM concentration and unlabelled ligand was added. Unlabelled N-terminal TFII-I was added to $^{15}N$-labelled RhoGAPFF1 to a final molar ratio 9:1 and to $^{15}N$-labelled RhoGAPFF4 to a final ratio 5:1. For the temperature series $^{15}N$-HSQC experiments were recorded for $^{15}N$-labelled samples of the RhoGAPFF1 domain and $^{15}N$-labelled CA150FF1 domain, in steps of 5K, ranging from 280K to 310K.

*Structure calculation*-Distance restraints derived from fully assigned peaks in NOESY experiments, $^{3}J(H^{N}, H^{\alpha})$ and hydrogen bond restraints were used for structure calculation. The

structures were calculated with the program CNS (Brünger et al, 1998) and ARIA 2.0 (Habeck et al, 2004), using 7 iterations of 20 structures and a final iteration of 80 structures. Water refinement was applied to the 30 lowest energy structures of the final iteration. The 15 lowest energy structures were analyzed with PROCHECK-NMR (Laskowski et al, 1996) and the statistics from the analysis are shown in Table 6. MOLMOL (Koradi et al, 1996) and PyMOL (DeLano, 2002) were used to visualize the structures and generate the figures.

*In vitro phosphorylation assays*-For the kinase assays, a purified RhoGAPFF1 domain sample and the FLT3 derived peptide DNEYFYV (used as a positive control) were tested. Assays were performed in 50 μl of 10mM HEPES (pH 7.5), 50mM NaCl, 5mM $MgCl_2$, 5mM $MnCl_2$, 1.25mM DTT, 0.2 mM ATP, 100 ng of PDGF-receptor α kinase (Cell Signaling Technology®) and either 3 μg of protein or the peptide at a final concentration of 3 μM. Samples were collected at 30ºC or 37ºC at a range of times. They were then concentrated and analyzed by MALDI-TOF.

The phosphorylation site within the RhoGAPFF1 domain was determined after tryptic digestion of the domain, followed by MALDI identification and MS/MS fragmentation of the phosphorylated peptide. In this case, the phosphorylation reaction was done with a 5-fold excess of kinase and left overnight at 37ºC, in order to obtain high amounts of phosphorylated domain for the analysis.

*Deposition of assignments and coordinates*

The $^{1}H$, $^{15}N$ and $^{13}C$ chemical shift assignments have been deposited in the BioMagResBank database (http://www.bmrb.wisc.edu) under the accession number 15938, and the protein coordinates in the protein data bank under accession number 2k85 (15 structures).

## 6.3 RESULTS

*Description of p190-A RhoGAPFF1 solution structure*

We used standard multidimensional heteronuclear NMR spectroscopy to assign the backbone and side-chain resonances of almost all residues in the p190-A RhoGAPFF1 domain. Table 6 summarizes the NOE, dihedral and hydrogen bond restraints used for the structure calculation and the statistics corresponding to the analysis of the 15 lowest-energy conformers selected from a set of 80 calculated structures. The superimposition of these structures is shown in Fig. 32 (a).

**Table 6**: **Structural statistics for the 15 lowest energy conformers of p190-A RhoGAP FF1 domain.**

| **Restraints used for the calculation** <SA>**(a)** | |
|---|---|
| Intraresidual | 0 |
| Sequential ( \|i-j\| = 1) | 373 |
| Medium range ( 1< \|i-j\| ≤ 4) | 239 |
| Long range ( \|i-j\| > 4) | 298 |
| Unambiguous | All |
| Ambiguous | 0 |
| Dihedrals | 82 |
| Hydrogen bonds | 70 |
| All | 911 |
| Restraint per residue ratio | 13.2 |
| **R.m.s. deviation (Å) from experimental (b)** | |
| NOE: | $0.001437 \pm 1.5 \times 10^{-4}$ |
| Bonds (Å) | $0.003117 \pm 1.5 \times 10^{-4}$ |
| Angles (º) | 0.4673 ± 0.02 |
| **Coordinate Precision (Å) (c)** | |
| backbone secondary structure elements | 0.24 |
| heavy atoms in secondary structure elements | 0.68 |
| heavy atoms all residues | 0.96 |
| **CNS potential energy (kcal $mol^{-1}$)** | |
| Total energy(d) | -2295.7 ± 46.2 |
| Electrostatic | -2332.4 ± 35.7 |
| van der Waals | -693.1± 6.5 |
| Bonds | 81.5 ± 6.1 |
| Angles | 171.2 ± 5.3 |
| **Structural quality (%residues)** 15 best structures | |
| in most favored region of Ramachandran plot | 84.7 |
| in additionally allowed region | 13.2 |

a) <SA> refers to the ensemble of the 15 structures with the lowest energy

b) No distance restraint in any of the structures included in the ensemble was violated by more than 0.3Å

c) r.m.s. deviation between the ensemble of structures <SA> and the lowest energy structure

d) $E_{L-J}$ is the Lennard-Jones van der Waals energy calculated using the CHARMM-PARMALLH6 parameters. $E_{L-J}$ was not included in the target function during the structure calculation.

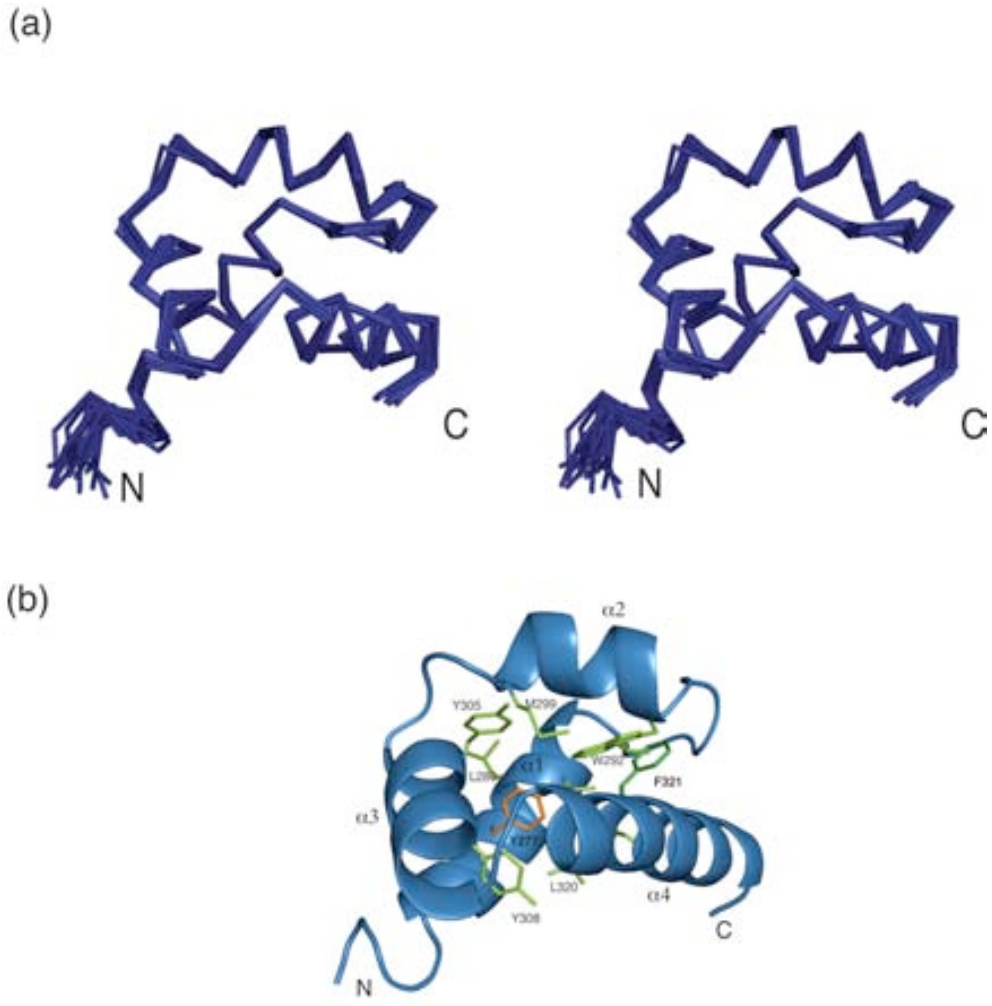



**Figure 32: Solution structure of the RhoGAPFF1 domain.**

**(a)** Stereo view of the backbone superimposition 15 lowest energy conformers of the RhoGAPFF1 domain after water refinement.

**(b)** Ribbon representation of the RhoGAPFF1 domain. The four helices span the following residues: Thr272-Ile284, Trp292-Met299, Glu304-Glu312 and Thr314-Lys329, respectively. Some of the semi-conserved hydrophobic residues that form the core of the domain are shown in green. The Tyr that replaces one of the characteristic conserved Phe is shown in orange, and the other conserved Phe, in dark green.

The RhoGAPFF1 domain is an all-alpha domain, formed by four α helices (on the basis of the hydrogen bond pattern of their helices), instead of the classic FF domain ($\alpha1-\alpha2-3_{10}-\alpha3$)

architecture observed so far in all available structures (Bonet et al, 2008). Fig. 32 (b) shows a cartoon representation of the secondary structural elements corresponding to RhoGAPFF1 and Fig. 33 (a) a comparison of this structure to that of FBP11FF1. The superimposition indicates that the two helices belong to distinct classes (the $3_{10}$ helix of FBP11FF1 displayed in red, the corresponding alpha helix of RhoGAPFF1 in yellow). The α3 helix present in RhoGAPFF1 is a turn longer than the corresponding classical $3_{10}$ helix found in other FF domains (marked in yellow on top of the alignment, Fig. 33 (b)).

All helices are packed against one another displaying a network of NOEs between semi-conserved residues, which define the hydrophobic core of the protein. These residues are shaded in green in the alignment of Fig. 33 (b) and some of them are depicted in Fig. 32 (b).

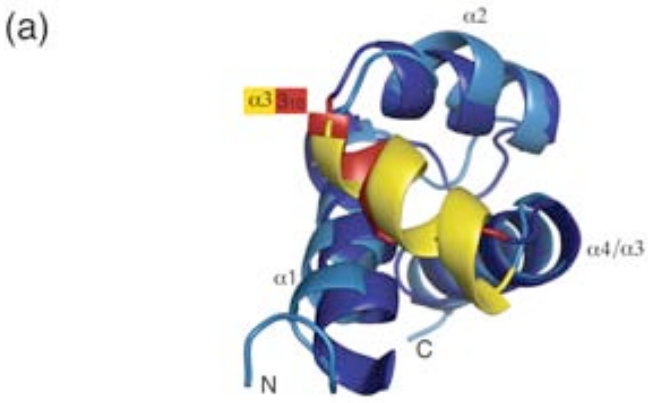



**Figure 33: Comparison of RhoGAPFF1 with other FF domains.**

**(a)** Superimposition of the typical $3_{10}$ helix (in red) of FBP11FF1 domain (in dark blue) and the extended α helix (in yellow) that replaces the $3_{10}$ helix in RhoGAPFF1 domain (in sky blue). The residues that form part of this extended α helix are marked in a yellow box in the alignment of Fig. 30 (b)

**(b)** Sequence alignment of FF domains from human p190-A RhoGAP, CA150 and FBP11. The alignment was generated with ClustalX (Thompson, 1997) and edited manually. The S symbol in a red box indicate the FF domains whose structure is deposited in the PDB (Berman et al, 2000). Conserved and semi-

conserved residues that form the hydrophobic core of the protein are indicated in green. Orange boxes mark the positions where one of the highly conserved Phe is substituted by another residue.

A second particular feature of the RhoGAPFF1 domain resides in the long loop connecting α1 and α2 helices. In all FF structures, including that of RhoGAPFF1, this loop is highly structured, since several residues are involved in contacts with the C-terminal end of the domain. Nevertheless, in RhoGAPFF1 these contacts to the C-terminal include Leu328 and Lys329, residues lying outside the general boundaries of FF domains. To elucidate the importance of these additional residues in the RhoGAPFF1 structure, we prepared a construct lacking the last four residues, thus adjusting the domain length to the classical FF boundaries. This shorter construct showed significantly poorer expression and indeed, the domain produced was unfolded (data not shown).

Interestingly, the first of the two characteristics conserved Phe is replaced by a Tyr (displayed in orange in Fig. 32 (b)) in the RhoGAPFF1 domain. To date, the RhoGAPFF1 domain constitutes the only solved structure of a FF domain that lacks one of the highly conserved Phe residues (Fig. 33 (b)). Therefore, although these two Phe are strongly conserved among FF domains, their presence is not indispensable for the correct folding of the domain. Furthermore, from the alignment of Fig. 33 (b), we observe that all the predicted FF domains of p190-A RhoGAP have, at least, one of the conserved Phe replaced by other residues.

In summary, the structure of the RhoGAPFF1 domain has some unexpected features, for instance, a new architecture including an insertion of three residues in the α3 helix, and the requirement of an additional turn in the last helix to obtain a folded sample.

*Tyr phosphorylation on the RhoGAPFF1 domain*

The solution structure of the RhoGAPFF1 domain revealed that the consensus site for the PDGF-receptor α mediated phosphorylation is located in the α3 helix and that Tyr308, the previously reported phosphorylation target (Jiang et al, 2005), is a conserved residue that participates in the formation of the hydrophobic core. Thus, in the structured domain the tyrosine hydroxyl group would be inaccessible to the kinase. Furthermore, the presence of a phosphate group at this position would compromise the stability of the domain. This hypothesis was confirmed with the production of a Y308D mutant, described to function as phospho-mimicking since it also exhibited a greatly reduced affinity for the interaction with the TFII-I (Jiang et al, 2005). The mutant was largely insoluble and precipitated after removing the fusion tag. This observation suggests that the RhoGAPFF1 fold is not compatible with the presence of a charged residue at position 308 (Fig. 34 (a)).

Further, the structure of RhoGAPFF1 showed that, in contrast to Tyr308, Tyr310, which is located in the C-terminal end of the helix α3, is solvent-exposed. Therefore, if Tyr phosphorylation takes place on the RhoGAPFF1 domain, Tyr310 is more likely to be the phosphorylation site, since phosphorylation at this position would not imply, in principle, the disruption of the FF structure (Fig 34 (b)).

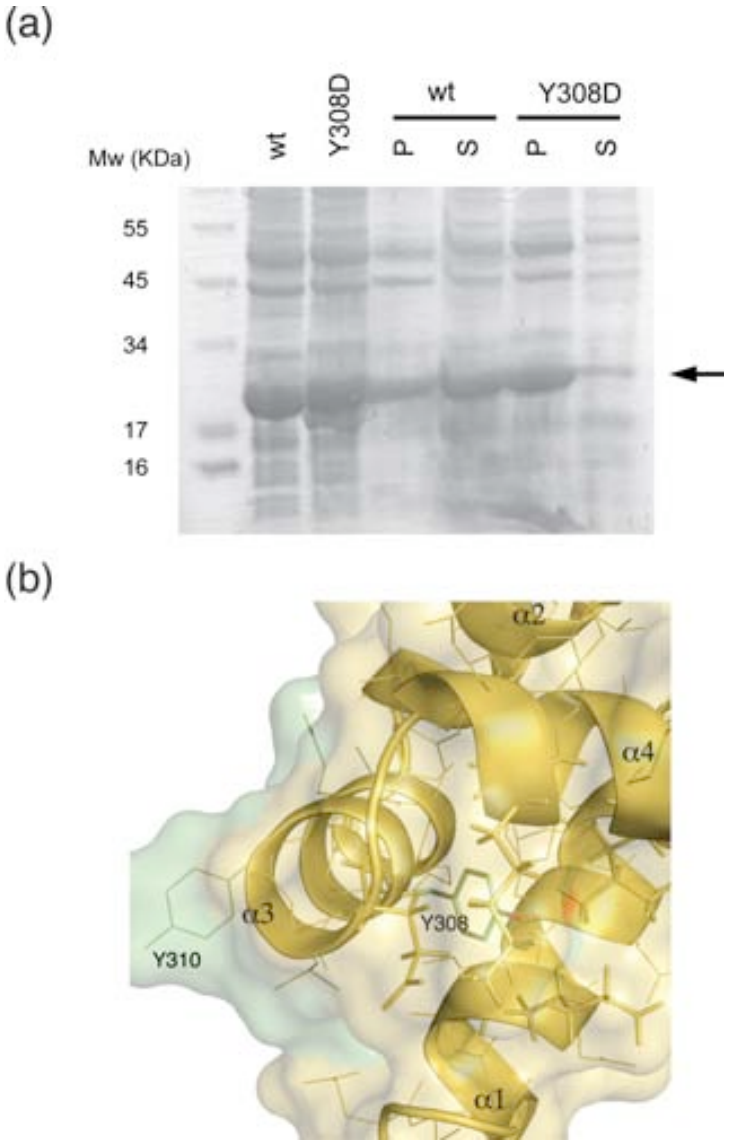



**Figure 34:**
**Y308 is a buried residue in RhoGAPFF1 and forms part of the hydrophobic core.**
**(a)** SDS-PAGE electrophoresis showing the different solubility of the *wt* domain and the phospho-mimicking (Y308D) mutant. *wt* and Y308D lanes correspond to total fractions, and P and S correspond to pellet and supernatant fractions, respectively. The arrow marks the band position for the RhoGAPFF1 domain.

**(b)** Detailed view of Tyr308 showing its buried side-chain, with the hydroxyl group marked in red, and Tyr310 showing its solvent-exposed accessible side-chain.

Thus, we examined the feasibility of the phosphorylation reaction and sought to identify the specific target for the kinase. The phosporylation assays on the RhoGAPFF1 domain were carried out using the FLT3-derived peptide DNEYFYV as a positive control, and the reaction was monitored by mass spectrometry. After two hours of reaction at 30ºC, hardly any sign of phosphorylation in the RhoGAPFF1 domain was detected, while in the FTL3 peptide the most abundant form was the phosphorylated one. However, leaving the kinase reaction for 5

additional hours at 37ºC yielded a peak corresponding to the phosphorylated form of the domain (Fig. 35).

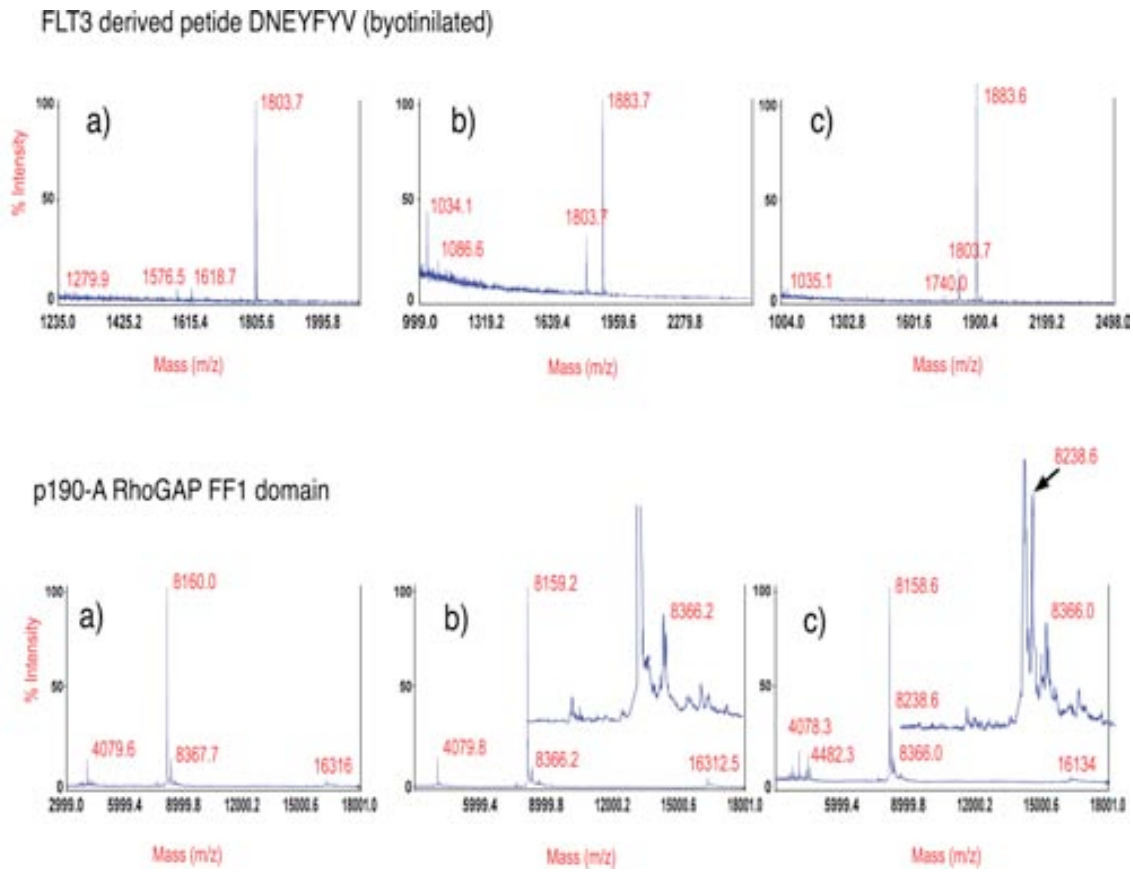



**Figure 35: MALDI-TOF spectra of the phosphorylation assays for the RhoGAPFF1 domain (below) and for the positive control peptide (above).**

Samples for the analysis were extracted before starting the reaction (a), after 2h at 30ºC (b) and after 5h at 37ºC (c). For the RhoGAPFF1 spectra (b) and (c), the region of the protein peaks has been zoomed, to better appreciate the appearance of the peak corresponding to the phosphorylated domain (marked with an arrow in (c)).

Furthermore, the amino acid fragmentation of the 18-mer phosphorylated peptide (generated by tryptic digestion of the phosphorylated FF domain) confirmed that Tyr308 is the main phosphorylation target of the PDGF-receptor α kinase within the RhoGAPFF1 domain (Fig 36).

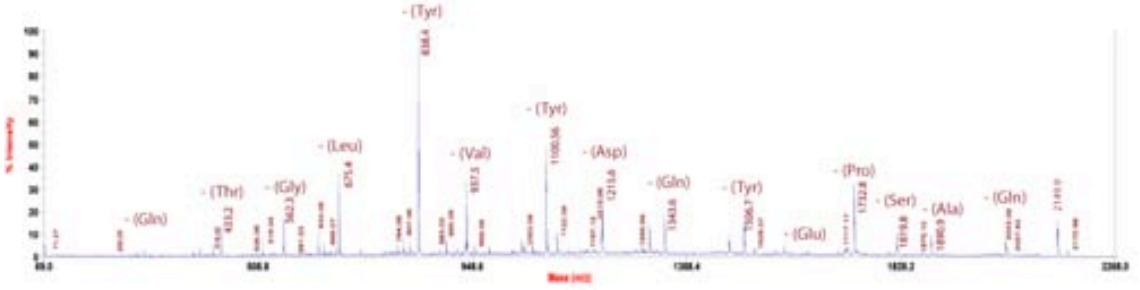
To characterize the conformation of the phosphorylated state of RhoGAPFF1 we scaled up the assays to obtain large amounts of the phosphorylated form but we observed that the domain substantially precipitated, indicating that the incorporation of a phosphate group at

Tyr308 causes a destabilization of the domain, in agreement with what we observed with the Y308D mutant.

MS/MS fragmentation of peptide:

| Res: | MQASPEYQDYVYLEGTQK (unphosphorylated) | MQASPEYQDpYVYLEGTQK (Y308 phosphorylated) | MQASPEYQDYVpYLEGTQK (Y310 phosphorylated) |
|---|---|---|---|
| Met 18 | - | - | - |
| Gln 17 | 2018.935 | 2098.901 | 2098.901 |
| Ala 16 | 1890.876 | 1970.842 | 1970.842 |
| Ser 15 | 1819.839 | 1899.805 | 1899.805 |
| Pro 14 | 1732.807 | 1812.773 | 1812.773 |
| Glu 13 | 1635.754 | 1715.721 | 1715.721 |
| Tyr 12 | 1506.712 | 1586.678 | 1586.678 |
| Gln 11 | 1343.648 | 1423.615 | 1423.615 |
| Asp 10 | 1215.590 | 1295.556 | 1295.556 |
| Tyr 9 | 1100.563 | 1180.529 | 1180.529 |
| Val 8 | 937.499 | 937.499 | 1017.466 |
| Tyr 7 | 838.431 | 838.431 | 918.397 |
| Leu 6 | 675.368 | 675.368 | 675.368 |
| Glu 5 | 562.284 | 562.284 | 562.284 |
| Gly 4 | 433.241 | 433.241 | 433.241 |
| Thr 3 | 376.220 | 376.220 | 376.220 |
| Gln 2 | 275.172 | 275.172 | 275.172 |
| Lys 1 | 147.113 | 147.113 | 147.113 |

RhoGAPFF1 derived peptide - unphosphorylated (Mw - 2149.9)




RhoGAPFF1 derived peptide - phosphorylated (Mw - 2230)

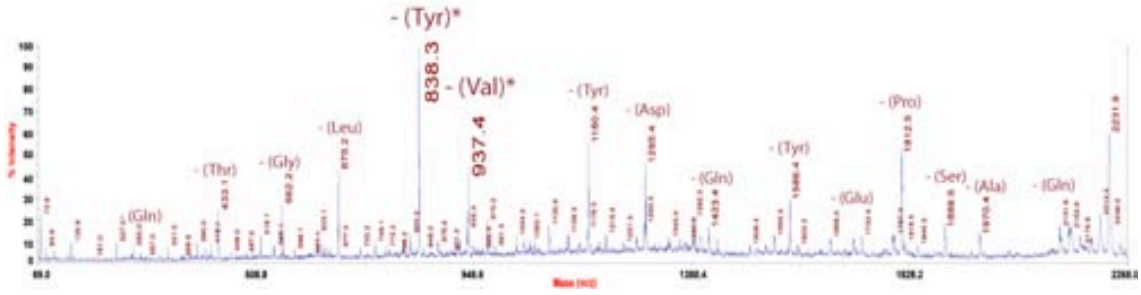



**Figure 36: Determination of the phosphorylation site within the RhoGAPFF1 domain.**
Above, a table is shown indicating the theoretical mass of the fragments generated for the unphosphorylated peptide and the two possible phosphorylated peptides (corresponding to Tyr308 and Tyr310). The distinct masses of the fragments highlighted in green and red allows discrimination between the two peptides. Below, the spectra of the fragmented unphosphorylated and phosphorylated peptides (with their masses) are shown. In the spectrum of the phosphorylated peptide, the peaks marked with an asterisk are those indicating that phosphorylation occurs at Tyr308.

These results confirm that RhoGAPFF1 is phosphorylated by the PDGF-receptor α kinase at Tyr308 (Jiang et al, 2005). More importantly, the phosphorylation appears to be temperature-dependent and to cause an irreversible destabilization of the RhoGAPFF1 domain.

*Dependence of RhoGAPFF1 domain folding/unfolding with temperature variations*

The temperature-dependent phosphorylation of the RhoGAPFF1 domain led us to test whether the RhoGAPFF1 structure is affected by variations in temperature. Consequently, we monitored its amide chemical shift changes induced by temperature as a set of $^{15}$N-HSQC experiments acquired at 280K, 285K, 295K 305K and 310K and we observed that peak dispersion and intensity is significantly reduced for the RhoGAPFF1 amide resonances (Fig. 37 (a) and (c)). Nevertheless, the process is reversible and RhoGAPFF1 peaks recover dispersion when the temperature is decreased from 310K to 280K. In parallel, we tested whether a classical FF structure behaved in a similar manner in response to temperature variations. In Fig. 37 (b) and (c) we show a similar set of experiments for CA150FF1 domain, which indicate that the domain is folded, with a good dispersion of their amide resonances along the temperature variations and that peak intensity augments with TºK. Only the N-terminal peaks of CA150FF1, which correspond to an unstructured region out of the boundaries of the FF domain, show a decreased intensity with TºK (Fig. 37 (c)). Thus, we can attribute the changes experienced by RhoGAPFF1 amide resonances to the appearance of unfolded populations of the domain within this range of TºK. In contrast, the CA150FF1 domain remains folded and stable.

Another indication of the high flexibility of RhoGAPFF1 domain arises from $D_2O$ exchange experiments (not shown). In these experiments, no peak could be detected in a $^{15}$N-HSQC experiment immediately after dissolving the domain in $D_2O$, indicating a rapid exchange of all the proton amides of the domain, even the ones in secondary structure elements that are supposed to have a slower rate of exchange due to their participation in hydrogen bond formation.

Overall, these observations would explain why RhoGAPFF1 is phosphorylated on Tyr 308 at 37ºC but not at low temperature.

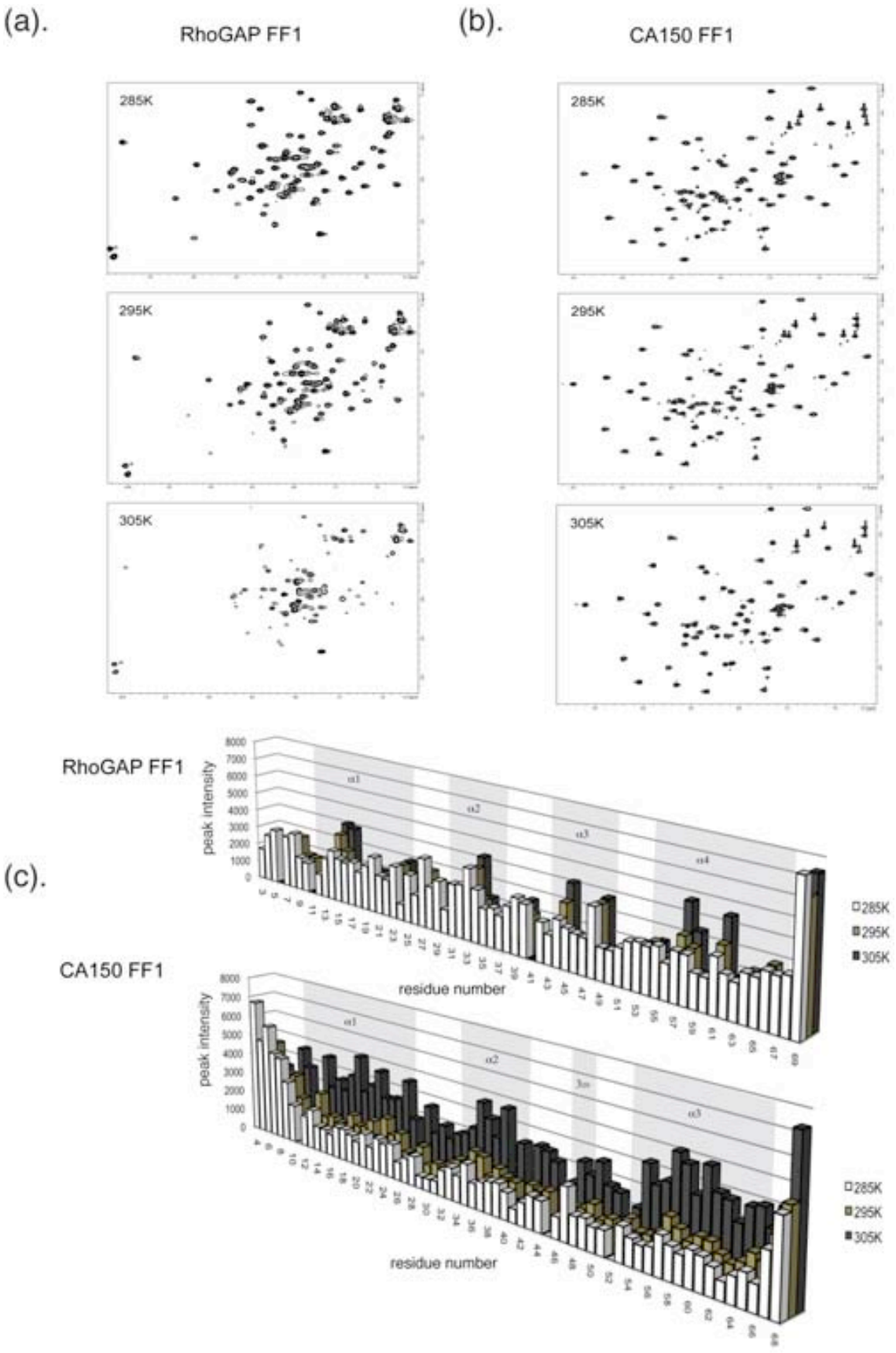



**Figure 37: Temperature effect in the stability of RhoGAPFF1 domain and CA150FF1 domain.**

**(a)** $^{15}$N-HSQC spectra at different TºK (285K-295K-305K) of RhoGAPFF1 domain that clearly show the change in the appearance of the peaks with the increase of the TºK.
**(b)** $^{15}$N-HSQC spectra at different TºK (285K-295K-305K) of CA150FF1 domain, that shows minimal differences in the appearance of the peaks with the increase of the TºK.
**(c)** Graphical representation of peak intensities as a function of the TºK for the RhoGAPFF1 and CA150FF1 domain. Peak intensity is represented for each residue at 285K-295K-305K. Intensity of peaks shows a general decrease at higher TºK in RhoGAPFF1 domain, whereas in CA150FF1 domain the intensity of peaks increases at higher TºK.

*The N-terminal fragment of TFII-I is unstructured in solution*

From a previous study on p190-A RhoGAP, it was reported that it interacted with the N-term part of the TFII-I transcription factor via its FF domains. Hence, we sought to examine the binding of p190-A RhoGAP FF domains to the TFII-I transcription factor. To test the interaction, we produced four different constructs of the TFII-I N-terminal, ranging from the entire N-terminal (amino acids 1-89) to a shorter one consisting of 40 residues (amino acids 21-60). All the constructs include the putative leucine zipper, but the deletion of the 30 C-terminal residues yielded highly insoluble proteins whereas the two constructs including the C-terminal were mostly soluble (Fig. 38 (a)). From these two fragments, only the longer construct including the first 20 residues was generated after protease cleavage.

As a result, we could only obtain large amounts of the longer construct of the N-terminal TFII-I. Remarkably, according to the purification profile in the gel filtration chromatography, the construct behaves as a dimer in solution, as previously observed in other studies (Roy, 2007), even when we performed the purification in presence of reducing agents (Fig 38 (b)). Analysis by mass spectrometry revealed two forms of the N-terminal TFII-I purified, the entire fragment (1-89) and a shorter one lacking 4 residues from the C-terminal. The analysis also confirmed the presence of a dimeric form of the N-terminal TFII-I.

Furthermore, the $^{15}$N-HSQC of the N-terminal TFII-I showed a spectrum with very few and poorly dispersed amide resonances. To favour the presence of monomeric forms of the N-terminal TFII-I, we recorded the experiments at different TºK up to 310K, but the overall aspect of the peaks did not change at all, suggesting that, in any case, the N-terminal TFII-I is unstructured in solution (Fig. 38 (c)).

*RhoGAPFF1 and FF4 domains do not bind the N-terminal of TFII-I transcription factor in an independent manner*

Although it was shown that TFII-I binding was optimal with a construct including the four FF domains of p190-A RhoGAP, each of the single domains also has the capacity to interact with the N-terminal of TFII-I (Jiang et al, 2005). Hence, in our studies, additionally to RhoGAPFF1, we examined the FF4 domain, since it has been reported to show the strongest interaction to TFII-I (Jiang et al, 2005).

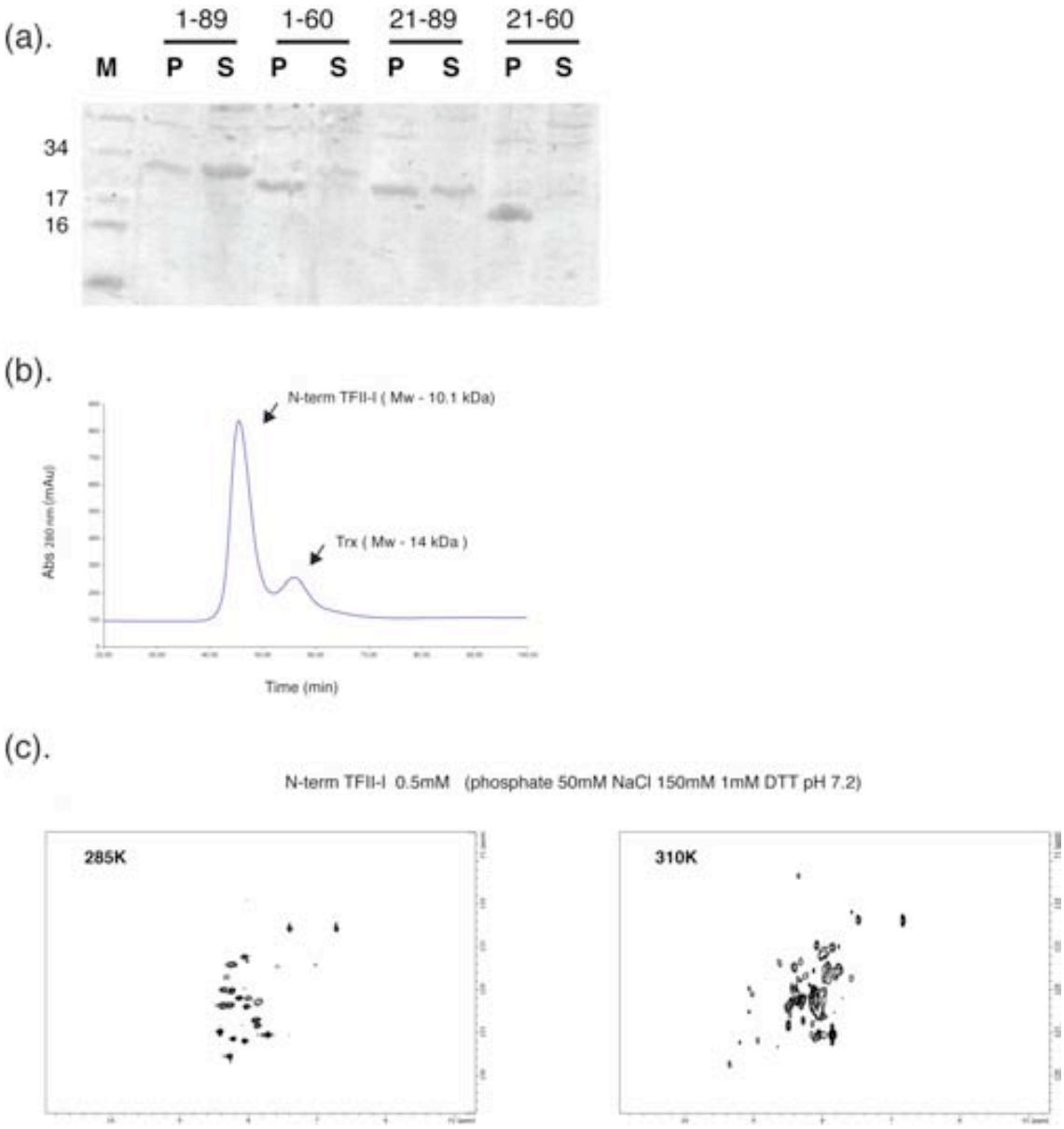



**Figure 38: The N-terminal fragment of TFII-I is unstructured in solution.**

**(a)** SDS-PAGE electrophoresis of pellet (P) and supernatant (S) fractions for the four expressed constructs of N-terminal TFII-I.

**(b)** Gel-filtration elution profile of the N-terminal TFII-I (1-89) and the fusion protein thioredoxin (Trx). Chromatography was performed in phosphate buffer 50mM pH 7.2, NaCl 150mM and DTT 1mM on a HiLoad™ Superdex™ 30 column (GE healthcare Life Sciences).

**(c)** $^{15}$N-HSQC spectra of the N-terminal TFII-I (1-89) at 285K (*left*) and 310K (*right*).

We performed NMR titration experiments using either a $^{15}$N-labelled sample of RhoGAPFF1 with increasing amounts of unlabelled N-terminal TFII-I or the opposite combination. We recorded the experiments at 285K and 310K, to ensure that the possible binding did not depend on the flexibility of RhoGAPFF1 domain. However, in all the cases, we observed only small changes in the amides of the FF domain upon addition of a high excess of the ligand (Fig. 39). In the same way, the addition of unlabelled RhoGAPFF1 to a $^{15}$N-labelled N-terminal TFII-I also did not result in changes on the unstructured appearance of the latter one, confirming the absence of binding observed between the two proteins in our experimental conditions.

Similarly, we carried out the same set of experiments with the RhoGAPFF4 domain and the N-terminal of TFII-I but no significant changes in the domain that could indicate an interaction between the two domains were detected. Furthermore, the $^{15}$N-HSQC of RhoGAPFF4 revealed that the construct we used is completely unfolded at 285K and at higher temperatures (not shown).

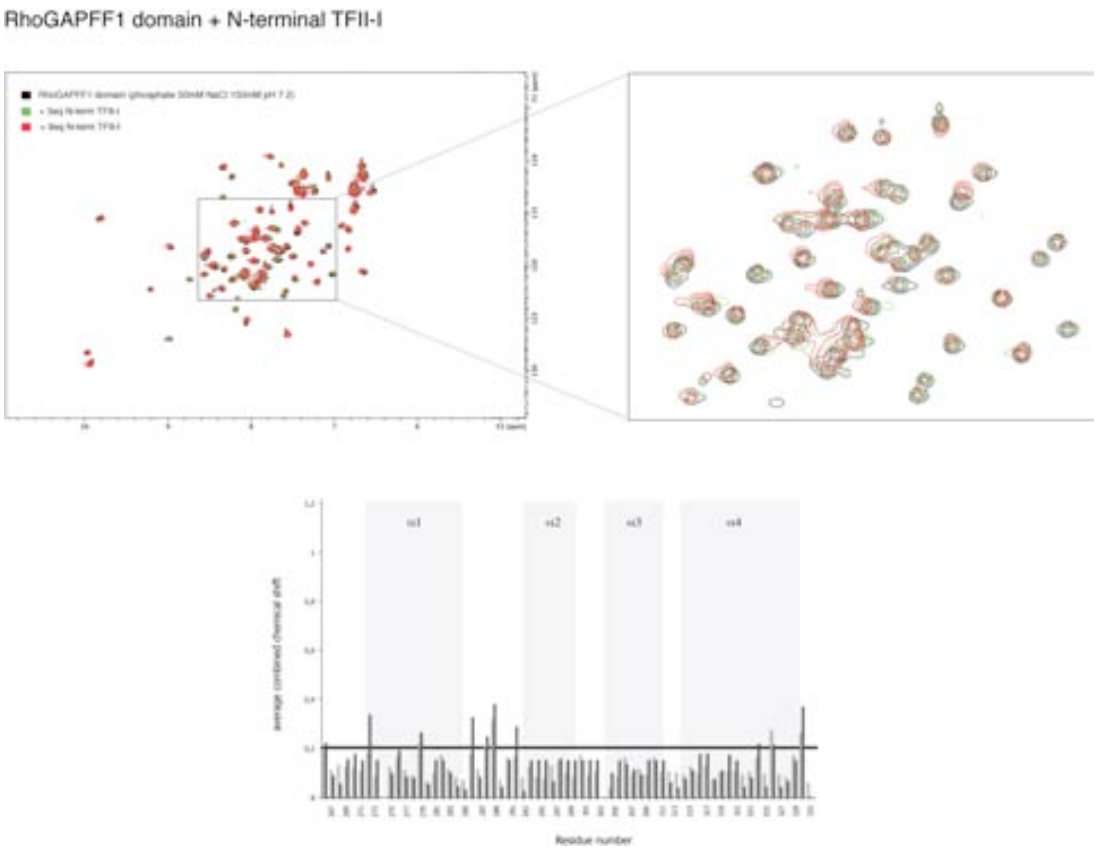



**Figure 39: Binding of the RhoGAPFF1 domain to the N-terminal fragment of TFII-I.**

*Up*- Overlay of the $^{15}$N-HSQC spectra of the $^{15}$N-labelled RhoGAPFF1 in absence (in *black*) and presence of increasing amounts of unlabelled N-terminal TFII-I (1-89). Molar ratios correspond to 1:5 *(green*) and

1:9 (*red*). The central region is zoomed to better appreciate the small changes on the RhoGAPFF1 amide resonances upon addition of high excess of ligand

*Down*- Bar diagram representation of the chemical shift changes for the RhoGAPFF1 upon addition of the N-term TFII-I (1-89). Light and dark blue bars represent changes at 1:5 and 1:9 ratios respectively. Chemical shifts were calculated from the equation $\Delta\delta^{av} = ((\Delta\delta_{1H})^2 + (\Delta\delta_{15N}/5)^2)^{1/2}$, where $\Delta\delta$ are the differences between a given value from the 1:5 or 1:9 spectrum and the equivalent value chemical shift from the reference spectrum.

In summary, we did not detect a significant interaction of either RhoGAPFF1 or FF4 with the N-terminal fragment of TFII-I under our experimental conditions.

## 6.4 DISCUSSION

FF domains are named based on the presence of two characteristic phenylalanine residues present in their sequence. The structure of the first FF domain of p190-A RhoGAP has the first Phe replaced by a Tyr, a conservative substitution that does not affect the overall structure.

Nevertheless, the structure of the RhoGAPFF1 domain displays several particular features compared to other FF structures previously described (Allen et al, 2002; Bonet et al, 2008; Gasch et al, 2006). First, it has a distinct architecture, with a $\alpha$3 helix instead of the classical $3_{10}$ helix present in other FF domains. Second, the RhoGAPFF1 domain contains an additional turn in the last helix, essential for the structure, since these residues participate in abundant interactions with the first loop. However, these additional contacts seem to be critical only for RhoGAPFF1 structure since the equivalent residues are absent in Prp40FF1 and URN1FF structures (Bonet et al, 2008; Gasch et al, 2006).

Another particularity of RhoGAPFF1 domain is that it progressively unfolds within a short range of temperatures - from 285K to 305K - in a reversible manner, a feature not observed in the CA150FF1 domain. Moreover, the ensemble of conformations that the RhoGAPFF1 domain exhibits in response to variations in temperature is crucial for the capacity of the domain to become phosphorylated by the PDGF-receptor $\alpha$. Furthermore, this temperature-dependent phosphorylation is explained by the confirmation of Tyr308 as the main phosphorylation site within the RhoGAPFF1 domain. As Tyr308 forms part of the hydrophobic core and thus is a buried residue in the structure, the domain should be unstructured to allow accessibility of Tyr308 to the kinase.

Moreover, a phosphate group at position Tyr308 interferes with the folding of the domain, as judged by the little solubility of the phospho-mimicking Y308D construct and the high degree of precipitation of RhoGAPFF1 after scaling up the phosphorylation assays. Therefore, the proposed inhibition of the interaction between the RhoGAPFF1 and the N-

terminal of TFII-I by phosphorylation could be explained by the stabilization of the unfolded/disordered state of the FF domain upon phosphorylation.

Under our experimental conditions we did not detect a significant interaction of either RhoGAPFF1 or RhoGAPFF4 with the N-terminal fragment of TFII-I. Indeed it has been suggested that the interaction between these two proteins is more efficient when all four FF domains are present and also when phosphatase inhibitors are added (Jiang et al, 2005). The first observation points to the possibility that FF domains act cooperatively. This mechanism would be in contrast, however, with another study that reported that the FF domains of CA150 possessed and independent binding activity from each other (Smith et al, 2004). Besides, the first FF domain of p190-A RhoGAP is spaced from the other FFs by a linker of 40 amino acids, so it is unlikely that it acts coordinately with the other FF domains to bind to ligands.

The second possibility implies that phosphorylation on TFII-I could be important for the binding. Perhaps, having the N-terminal TFII-I in a phosphorylated state at certain positions would favour the binding to p190-A RhoGAP FF domains, analogously to the case of the binding of CA150 FF domains to the C-terminal repeats (CTD) of RNApol-II (Carty et al, 2000).

In summary, our study has revealed new structural features and a particular mechanism for the PDGF-receptor α mediated phosphorylation of the RhoGAPFF1 domain. However, the role of this mechanism as a way to disrupt the p190-A RhoGAP FFs interaction to the N-terminal TFII-I is unclear, since if that binding takes place, it should be upon previous phosphorylation events on the N-terminal TFII-I. Curiously, in the binding between the N-terminal TFII-I and the Btk kinase, phosphorylation on a Tyr within the TFII-I disrupts the interaction and restores the dimeric state of the N-terminal TFII-I (Sacristan et al, 2004). Contrarily, in our proposal, phosphorylation on the N-terminal TFII-I would have the opposite function, that is, break the dimeric forms to allow the interaction with the FF domains of p190-A RhoGAP. In this context, the phosphorylation on Tyr308 within the RhoGAPFF1 domain could have a role in disrupting the interaction.

This hypothesis, however, should be addressed in further experiments to confirm the existence of phosphorylation sites within the N-terminal TFII-I and the implication for the binding with the FF domains of p190-A RhoGAP.

# GENERAL DISCUSSION

FF domains were first identified in 1999 by Bedford M. and Leder P., from sequence-database searches among splicing-related proteins. Concretely, they found that the WW-containing proteins FBP11 and CA150 display sequence homology in C-terminal regions respect to the WW domain. This newly discovered motif was termed FF because of the observation that they contained two highly conserved phenylalanine residues, and it was described as a novel motif that often accompanied WW domains (Bedford & Leder, 1999). FF domains are present in three protein families: the splicing factors FBP11, Prp40 and URN1, the transcription factor CA150 and the cytoplasmatic p190 RhoGTPase activating proteins (GAPs). This simplicity in distribution, however, is contrasted by the difficulty in defining their biological role.

Like WW domains, FF have been defined as protein interaction domains because the first specific function reported for them was the association of CA150 FF domains to the phosphorylated C-terminal domain of RNA-polymerase II (Carty et al, 2000). These findings also lead to the definition of the FF domain as a phospho-serine binding module. Nevertheless, this is not an accurate definition since it has been observed that some FF domains show ability to interact with non-phosphorylated peptides (Ester & Uetz, 2008; Gasch et al, 2006). In fact, the group of ligands that have been described for FF domains include a variety of factors, with binding reports pointing, in addition to the referred phospho-CTD, to negative/aromatic sequences, a tetratricopeptide repeat from the splicing factor Clf1, the transcription factor TFII-I and even to RNA. Table 7, extracted from (Ester & Uetz, 2008) summarizes all the found ligands for distinct FF domains.

| FF- containing protein | FF domain | Target protein | Binding region |
|---|---|---|---|
| Prp40 | regions including WW / FF | phospho-CTD of RNA-pol II | YpSPTpSPSYpSPTpSPS |
| | FF1 | Clf1 | TPR motif (amino acids 31-64) |
| | FF1 | Luc7 | FLGKIHLG |
| | FF2-3 | Snu71 | NDVHY |
| CA150 | FF1-5 | phospho-CTD of RNA-pol II | YpSPTpSPSYpSPTpSPS |
| | FF1-3 | Tat-Sf1 | $(D/E)_{2/5}$ - F/W/Y - $(D/E)_{2/5}$ |
| FBP11 | FF1 | phospho-CTD of RNA-pol II | YpSPTpSPSYpSPTpSPS |
| p190-A RhoGAP | FF1-4 | TFII-I | N-terminal fragment (1-90) |

**Table 7: Interacting partners reported for distinct FF domains.**

The table illustrates the difficulty to establish a consensus binding site for the FF domains. Moreover, it appears that FF domains are equally capable to interact with linear peptide motifs or to mediate domain-domain interactions.

Furthermore, unlike many protein interaction domains, including SH3, SH2, PDZ or WW, which have been widely studied with a high number of three-dimensional structures solved, for FF domains the structural information available is still limited. In fact, to date, there are only two reported structures of FF domains, which correspond to the first FF domain of FBP11 and to the first FF domain of Prp40 (Allen et al, 2002; Gasch et al, 2006).

In the first part of this thesis, we selected the only FF domain from yeast URN1 splicing factor as the subject for structural studies to expand our knowledge on the FF domain. The URN1 protein is one of the two known proteins containing only one FF domain, making it the most simplified representative of FF domain-containing splicing factors. The structure of the URN1FF is similar to that of other FF domains described, with the classic $\alpha1$- $\alpha2$-$3_{10}$- $\alpha3$ fold. However, URN1FF presents a distinctive negatively charged patch on its surface. Indeed, all available FF structures have a well-conserved fold but variable electrostatic patches on their surfaces. We observed that these patches are poorly conserved even for domains with similar overall $pK_a$ values. $pK_a$ values had been proposed to be a good indicator of FF binding specificity (Gasch et al, 2006; Smith et al, 2004) since FF displaying basic $pK_a$ values (for instance, FBP11FF1) could interact with the phospho-CTD whereas the FF1 of Prp40, with an acidic overall $pK_a$ did not show ability to bind to the phospho-CTD. Nevertheless, we showed that URN1FF is unable to bind to the reported ligand for the Prp40FF1 (the TPR motif of Clf1), indicating that both FF domains possess different specificities, albeit having a similar $pK_a$ value. Consequently, $pK_a$ value may not always be directly related to conservation of charge distribution on the surface, which is more related to sequence conservation. Hence, the overall $pK_a$ value of a given FF domain appears to be an inaccurate indicator of the binding specificity.
To investigate potential binding sites in FF domains, we performed structural comparisons to other proteins with similar folds. In addition to the structures grouped with FF domains in the SCOP database, we included another family of protein interaction domains, the SURP domains, because we demonstrated that their fold is also similar to that of FF domains, with the same architecture $\alpha1$- $\alpha2$-$3_{10}$- $\alpha3$ even though there is no sequence conservation between the two domain families. We observed that the main difference between all these structures resides in the orientation of the $\alpha2$ and $3_{10}$ helices. Remarkably, in DEK, SF3a120SURP2 and Prp40FF1 structures the $\alpha2$ helix participates in ligand recognition (Devany et al, 2004; Gasch et al, 2006; Kuwasako et al, 2006). The only exception is the binding of the FBP11FF1 to the phospho-CTD, where only the N-terminal parts of $\alpha1$ and $\alpha3$ participate in the interaction (Allen et al, 2002).

The results presented in the second part of this thesis support the implication of $\alpha2$ in ligand recognition. Indeed, we observed that the binding site of the CA150 FF domains for the

interaction with the phospho-CTD mainly involve residues in the first loop, the α2 and the $3_{10}$ helices. Consequently, the interacting region within the CA150 FF domains is different from that of FBP11FF1 for the same ligand. Thus, it seems that FF domains possess the ability to interact with the phospho-CTD repeats using different binding sites. Furthermore, in the case of FBP11FF1, the interacting surface included a cluster of positively charged residues that emphasized the importance of electrostatic contacts for the binding. However, in the case of CA150 FF domains, regions including hydrophobic patches are involved in the interaction, as observed for Prp40FF1 and SF3a120SURP2 binding sites. Thus, the FF fold seems to have developed binding site variations to accommodate an abundant and variable set of ligands.

In the last part of the thesis, we focused in another interesting feature of some of the FF domains, that is, the role that protein phosphorylation seems to play in regulating its interactions. Moreover, this regulation appears to be possible in the two directions, either by phosphorylation of the ligand or the domain. Examples of the first case - interaction of human CA150, FBP11 and yeast Prp40 FF domains to the phospho-CTD - have already been discussed in the previous sections of this work. As an example of the second case we find the interaction of the FF domains of human p190-A RhoGTPase activating protein (GAP) – which are, together with p190-B, the only cytoplasmatic proteins containing FF domains – with the general transcription factor TFII-I, which has been shown to be modulated by phosphorylation within the first FF domain of p190-A RhoGAP (Jiang et al, 2005). Nevertheless, the structural determinants governing this interaction remain unknown, so we decided to solve the structure of the RhoGAPFF1 domain, to identify the binding site for the interaction with the N-terminal fragment of TFII-I and to examine the impact of phosphorylation on the structure of the FF domain.

The solution structure of RhoGAPFF1 domain showed some interesting features. First, we found that the domain does not have the typical $3_{10}$ helix characteristic of FF domains. Instead, it presents a α1–α2–α3 –α4 topology. Also, we observed that contacts between the first loop of the structure and residues of the C-terminal end lying outside of the classical FF domain boundaries are indispensable for the correct folding and stability the domain. Curiously, in the structure of the FBP11FF1 domain, additional residues were also required for the correct folding of the domain, but they were situated at the N-terminal of the domain.

More importantly, the structure revealed that the residues that conform the phosphorylation site form part of the α3 helix and that the previously reported phosphorylation target Tyr308 is a semi-conserved residue that forms part of the hydrophobic core. Thus it appeared to be unlikely that phosphorylation could occur at this position because, generally, phosphorylation processes require disordered regions in protein targets in order to facilitate accessibility of the kinase to the binding site (Iakoucheva et al, 2004). In fact, we observed that only when the domain is

unstructured it could be phosphorylated. Therefore, phosphorylation requires the previous domain unfolding in a process that for RhoGAPFF1 occurs within a short range of temperatures (at 37ºC the domain is already largely disordered) compared to other FF domains that we studied such as CA150FF1 and FF2.

However, what remains unclear from our work is the role of FF1 phosphorylation in disrupting the interaction of p190-A RhoGAP FF domains with the TFII-I, since we observed that, in our conditions, the binding of FF1 and FF4 domains to the N-terminal fragment of TFII-I is very weak. As discussed earlier, it might be that serine/threonine phosphorylation within the ligand is important for an efficient interaction, and this hypothesis would also point to the preference that FF domains have for phosphorylated ligands.

This lack of binding detected in our experiments for the FF domains from p190-A RhoGAP and the N-terminal fragment of TFII-I evidences the difficulty in defining the specificity of ligand recognition for the FF domains that we illustrated in Table 7. Another case of an unclear association would be that of Prp40 domains with the phospho-CTD. In addition to the previously reported inability of Prp40 WW domains to bind the phospho-CTD (Wiesner et al, 2002), we did not detected binding for the FF1-2 pair and the FF4 domains of Prp40 to the phospho-CTD. Hence, if there is binding of Prp40 domains to the phospho-CTD it should be mediated by a mechanism in which the presence of both classes of domains enhances the interaction.

On the contrary, we have confirmed the interaction between CA150 FF domains and the phospho-CTD although in our hands, and in contrast to what was reported in previous studies (Carty et al, 2000), all the domains possessed the same ability to bind the phospho-CTD, with roughly an equal weak affinity on the basis of NMR titration experiments. This low affinity detected for the FF associations could be one of the factors that have prevented to date the availability of complex structures for the FF domains with their ligand partners. Clearly, this kind of structural information together with the search and identification of new binding partners should be the aim of future work and would help in a further characterization of the FF domains.

# CONCLUSIONS

## Chapter 1

1. We determined the solution structure of URN1FF and Prp40FF4 domains. The two domains have the classical $\alpha1$- $\alpha2$-$3_{10}$- $\alpha3$ topology observed in the previously described FBP11FF1 and Prp40FF1 domains. However, the long loop connecting the first and the second alpha helices is extended in the Prp40FF4 structure compared to previously reported FF domains.

2. Despite low sequence identity, superimposition of distinct FF structures is good in all cases, with the most variable region comprising the first loop, the $\alpha2$ and the $3_{10}$ helices.

3. The surface charge distribution among FF structures is not conserved even for domains with similar overall $pK_a$ values. For instance, URN1FF and Prp40FF1 have similar $pK_a$ values (around 4.5), however URN1FF does not show interaction with the TPR motif reported to bind Prp40FF1, thereby confirming that $pK_a$ values are poor indicators of domain specificity.

4. The superimposition of URN1FF to the members of the same fold group (according to SCOP database) and to the SURP domains give similar values. Therefore, from a structural point of view, the SURP modules could be included in the FF fold.

5. The main difference between members of the FF fold (including FF domains) and SURP domains lies in the orientation of the second helix. Remarkably, except for the FBP11FF1 domain binding to the phospho-CTD, the $\alpha2$ helix participates in ligand recognition in all the reported interactions of the FF fold members with targets.

6. Neither the Prp40 FF4 nor the Prp40 FF1-2 pair showed ability to interact with the phospho-CTD.

## Chapter 2

1. In our experimental conditions all FF domains of CA150 interacted with the phospho-CTD with a similar low affinity. Binding to the phospho-CTD does not contribute to the folding of the intrinsically disordered FF5 and FF6 domains.

2. The binding sites within the FF1-4 domains are located in regions including the first loop, the α2 and the $3_{10}$ helices, and thus is different from the binding site described for the FBP11FF1 interaction with the same ligand.

3. As judged by HSQC experiments, the binding mode and affinity of the individual FF1 and FF2 domains is similar to that of a construct including both domains. Hence, despite being generally present as multiple copies in proteins, FF domains do not appear to exhibit cooperativity in binding to ligands.

## Chapter 3

1. We determined the solution structure of p190-A RhoGAP FF1 domain. Instead of the classical FF fold, it has a α1- α2- α3- α4 architecture. We showed that contacts between the first loop and a C-terminal extension are indispensable for the folding of the domain. In addition, we observed that Tyr308, the phosphorylation target for the RhoGAPFF1 is buried in the structure and participates in the hydrophobic core.

2. We showed that RhoGAPFF1 domain experiences a reversible unfolding within a short range of temperatures compared to other FF domains. Moreover, phosphorylation by the PDGF-receptor α only occurs when the domain is unstructured, and consequently the phosphorylation site accessible.

3. The presence of a charged group at position 308 seems to be incompatible with the FF fold, as indicated by the little solubility of the Y308D mutant and by the irreversible destabilization that phosphorylation causes on the RhoGAPFF1 domain.

4. In our conditions we did not detect binding of the p190-A RhoGAP FF1 and FF4 domains to the N-terminal fragment of TFII-I. Thus, it seems that phosphorylation within the N-terminal TFII-I sequence could be important for an efficient interaction.

# REFERENCES

Abovich N, Rosbach M (1997) Cross-intron bridging interactions in the yeast commitment complex are conserved in mammals. *Cell* **89:** 403-412

Allen M, Friedler A, Schon O, Bycroft M (2002) The structure of an FF domain from human HYPA/FBP11. *J Mol Biol* **323**(3)**:** 411-416

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**(3)**:** 403-410

Arthur WT, Burridge K (2001) RhoA inactivation by p190RhoGAP regulates cell spreading and migration by promoting membrane protrusion and polarity. *Mol Biol Cell* **12**(9)**:** 2711-2720

Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A* **98**(18)**:** 10037-10041

Bartels C, Xia T-H, Billeter M, Güntert P, Wüthrich K (1995) The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J Biomol NMR* **5:** 1-10

Bedford MT, Leder P (1999) The FF domain: a novel motif that often accompanies WW domains. *Trends Biochem Sci* **24**(7)**:** 264-265.

Bedford MT, Reed R, Leder P (1998) WW domain-mediated interactions reveal a spliceosome-associated protein that binds a third class of proline-rich motif: The proline glycine and methionine-rich motif. *Proc Natl Acad Sci* **95:** 10602-10607

Berman HM, Bhat TN, Bourne PE, Feng Z, Gilliland G, Weissig H, Westbrook J (2000) The Protein Data Bank and the challenge of structural genomics. *Nat Struct Biol* **7 Suppl:** 957-959

Bhattacharyya RP, Remenyi A, Yeh BJ, Lim WA (2006) Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits. *Annu Rev Biochem* **75:** 655-680

Bonet R, Ramirez-Espain X, Macias MJ (2008) Solution structure of the yeast URN1 splicing factor FF domain: comparative analysis of charge distributions in FF domain structures-FFs and SURPs, two domains with a similar fold. *Proteins* **73**(4)**:** 1001-1009

Bradley WD, Hernandez SE, Settleman J, Koleske AJ (2006) Integrin signaling through Arg activates p190RhoGAP by promoting its binding to p120RasGAP and recruitment to the membrane. *Mol Biol Cell* **17**(11)**:** 4827-4836

Brünger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL (1998) Crystallography and NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D* **54:** 905-921

Bustelo XR, Sauzeau V, Berenjeno IM (2007) GTP-binding proteins of the Rho/Rac family: regulation, effectors and functions in vivo. *Bioessays* **29**(4)**:** 356-370

Carty SM, Goldstrohm AC, Sune C, Garcia-Blanco MA, Greenleaf AL (2000) Protein-interaction modules that organize nuclear function: FF domains of CA150 bind the phosphoCTD of RNA polymerase II. *Proc Natl Acad Sci U S A* **97**(16)**:** 9015-9020

Castresana J, Saraste M (1995) Does Vav bind to F-actin through a CH domain? *FEBS Lett* **374**(2)**:** 149-151

Chan SP, Kao DI, Tsai WY, Cheng SC (2003) The Prp19p-associated complex in spliceosome activation. *Science* **302**(5643)**:** 279-282

Chang JH, Gill S, Settleman J, Parsons SJ (1995) c-Src regulates the simultaneous rearrangement of actin cytoskeleton, p190RhoGAP, and p120RasGAP following epidermal growth factor stimulation. *J Cell Biol* **130**(2)**:** 355-368

Chen CH, Yu WC, Tsao TY, Wang LY, Chen HR, Lin JY, Tsai WY, Cheng SC (2002) Functional and physical interactions between components of the Prp19p-associated complex. *Nucleic Acids Res* **30**(4)**:** 1029-1037

Cheriyath V, Roy AL (2001) Structure-function analysis of TFII-I. Roles of the N-terminal end, basic region, and I-repeats. *J Biol Chem* **276**(11)**:** 8377-8383

Chung S, McLean MR, Rymond BC (1999) Yeast ortholog of the Drosophila crooked neck protein promotes spliceosome assembly through stable U4/U6.U5 snRNP addition. *RNA* **5**(8)**:** 1042-1054.

Civera C, Simon B, Stier G, Sattler M, Macias MJ (2005) Structure and dynamics of the human pleckstrin DEP domain: distinct molecular features of a novel DEP domain subfamily. *Proteins* **58**(2)**:** 354-366

D'Andrea LD, Regan L (2003) TPR proteins: the versatile helix. *Trends Biochem Sci* **28**(12)**:** 655-662

Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* **6:** 277-293

DeLano WL (2002) The PyMOL Molecular Graphics System.

Denhez F, Lafyatis R (1994) Conservation of regulated alternative splicing and identification of functional domains in vertebrate homologs to the Drosophila splicing regulator, suppressor-of-white-apricot. *J Biol Chem* **269**(23)**:** 16170-16179

Devany M, Kotharu NP, Matsuo H (2004) Solution NMR structure of the C-terminal domain of the human protein DEK. *Protein Sci* **13**(8)**:** 2252-2259

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

Dziembowski A, Ventura AP, Rutz B, Caspary F, Faux C, Halgand F, Laprevote O, Seraphin B (2004) Proteomic analysis identifies a new complex required for nuclear pre-mRNA retention and splicing. *EMBO J* **23**(24)**:** 4847-4856

Ellis C, Moran M, McCormick F, Pawson T (1990) Phosphorylation of GAP and GAP-associated proteins by transforming and mitogenic tyrosine kinases. *Nature* **343**(6256)**:** 377-381

Ester C, Uetz P (2008) The FF domains of yeast U1 snRNP protein Prp40 mediate interactions with Luc7 and Snu71. *BMC Biochem* **9:** 29

Etienne-Manneville S, Hall A (2002) RhoGTPases in cell biology. *Nature* **420**(6916)**:** 629-635

Feng S, Chen JK, Yu H, Simon JA, Schreiber SL (1994) Two binding orientations for peptides to the Src SH3 domain: development of a general model for SH3-ligand interactions. *Science* **266**(5188)**:** 1241-1247

Fernandez C, Wider G (2003) TROSY in NMR studies of the structure and function of large biological macromolecules. *Curr Opin Struct Biol* **13**(5)**:** 570-580

Fortes P, Bilbao-Cortes D, Fornerod M, Rigaut G, Raymond W, Seraphin B, Mattaj IW (1999) Luc7p, a novel yeast U1 snRNP protein with a role in 5' splice site recognition. *Genes Dev* **13**(18)**:** 2425-2438

Gasch A, Wiesner S, Martin-Malpartida P, Ramirez-Espain X, Ruiz L, Macias MJ (2006) The structure of Prp40 FF1 domain and its interaction with the crn-TPR1 motif of Clf1 gives a new insight into the binding mode of FF domains. *J Biol Chem* **281**(1)**:** 356-364

Goldstrohm AC, Albrecht TR, Sune C, Bedford MT, Garcia-Blanco MA (2001) The transcription factor CA150 interacts with RNA polymerase II and the pre-mRNA splicing factor SF1. *Mol Cell Biol* **21:** 7617-7628

Gottschalk A, Tang J, Puig O, Salgado J, Neubauer G, Colot HV, Mann M, Seraphin B, Rosbash M, Luhrmann R, Fabrizio P (1998) A comprehensive biochemical and genetic analysis of the yeast U1 snRNP reveals five novel proteins. *RNA* **4**(4)**:** 374-393

Grainger RJ, Beggs JD (2005) Prp8 protein: at the heart of the spliceosome. *RNA* **11**(5)**:** 533-557

Habeck M, Rieping W, Linge JP, Nilges M (2004) NOE assignment with ARIA 2.0: the nuts and bolts. *Methods Mol Biol* **278:** 379-402

Hagiwara M, Nojima T (2007) Cross-talks between transcription and post-transcriptional events within a 'mRNA factory'. *J Biochem* **142**(1)**:** 11-15

Heasman SJ, Ridley AJ (2008) Mammalian RhoGTPases: new insights into their functions from in vivo studies. *Nat Rev Mol Cell Biol* **9**(9)**:** 690-701

Hernandez SE, Settleman J, Koleske AJ (2004) Adhesion-dependent regulation of p190RhoGAP in the developing brain by the Abl-related gene tyrosine kinase. *Curr Biol* **14**(8)**:** 691-696

Hillier BJ, Christopherson KS, Prehoda KE, Bredt DS, Lim WA (1999) Unexpected modes of PDZ domain scaffolding revealed by structure of nNOS-syntrophin complex. *Science* **284**(5415)**:** 812-815

Hong EL BR, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Livstone MS, Nash R, Oughtred R, Park J, Skrzypek M, Starr B, Andrada R, Binkley G, Dong Q, Hitz BC, Miyasato S, Schroeder M, Weng S, Wong ED, Zhu KK, Dolinski K, Botstein D, and Cherry JM. "Saccharomyces Genome Database".

Hu KQ, Settleman J (1997) Tandem SH2 binding sites mediate the RasGAP-RhoGAP interaction: a conformational mechanism for SH3 domain regulation. *EMBO J* **16**(3)**:** 473-483

Hyvonen M, Macias MJ, Nilges M, Oschkinat H, Saraste M, Wilmanns M (1995) Structure of the binding site for inositol phosphates in a PH domain. *EMBO J* **14**(19)**:** 4676-4685

Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res* **32**(3)**:** 1037-1049

Jiang W, Sordella R, Chen GC, Hakre S, Roy AL, Settleman J (2005) An FF domain-dependent protein interaction mediates a signaling pathway for growth factor-induced gene expression. *Mol Cell* **17**(1)**:** 23-35

Kajander T, Cortajarena AL, Mochrie S, Regan L (2007) Structure and stability of designed TPR protein superhelices: unusual crystal packing and implications for natural TPR proteins. *Acta Crystallogr D Biol Crystallogr* **63**(Pt 7)**:** 800-811

Kao HY, Siliciano PG (1996) Identification of Prp40, a novel essential yeast splicing factor associated with the U1 small nuclear ribonucleoprotein particle. *Mol Cell Biol* **16**(3)**:** 960-967

Koradi R, Billeter M, Wüthrich K (1996) MOLMOL: a program for display and analysis macromolecular structures. *J Mol Graphics* **14:** 51-55

Kuwasako K, He F, Inoue M, Tanaka A, Sugano S, Guntert P, Muto Y, Yokoyama S (2006) Solution structures of the SURP domains and the subunit-assembly mechanism within the splicing factor SF3a complex in 17S U2 snRNP. *Structure* **14**(11)**:** 1677-1689

Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR* **8**(4)**:** 477-486

Liu Q, Berry D, Nash P, Pawson T, McGlade CJ, Li SS (2003) Structural basis for specific binding of the Gads SH3 domain to an RxxK motif-containing SLP-76 peptide: a novel mode of peptide recognition. *Mol Cell* **11**(2)**:** 471-481

Macias MJ, Musacchio A, Ponstingl H, Nilges M, Saraste M, Oschkinat H (1994) Structure of the pleckstrin homology domain from beta-spectrin. *Nature* **369**(6482)**:** 675-677

MacMillan AM, Query CC, Allerson CR, Chen S, Verdine GL, Sharp PA (1994) Dynamic association of proteins with the pre-mRNA branch region. *Genes Dev* **8**(24)**:** 3008-3020

Mammoto A, Connor KM, Mammoto T, Yung CW, Huh D, Aderman CM, Mostoslavsky G, Smith LE, Ingber DE (2009) A mechanosensitive transcriptional mechanism that controls angiogenesis. *Nature* **457**(7233)**:** 1103-1108

Marley J, Lu M, Bracken C (2001) A method for efficient isotopic labeling of recombinant proteins. *J Biomol NMR* **20**(1)**:** 71-75

Meinhart A, Kamenski T, Hoeppner S, Baumli S, Cramer P (2005) A structural perspective of CTD function. *Genes Dev* **19**(12)**:** 1401-1415

Morales B, Ramirez-Espain X, Shaw AZ, Martin-Malpartida P, Yraola F, Sanchez-Tillo E, Farrera C, Celada A, Royo M, Macias MJ (2007) NMR structural studies of the ItchWW3 domain reveal that phosphorylation at T30 inhibits the interaction with PPxY-containing ligands. *Structure* **15**(4)**:** 473-483

Moran MF, Polakis P, McCormick F, Pawson T, Ellis C (1991) Protein-tyrosine kinases regulate the phosphorylation, protein interactions, subcellular distribution, and activity of p21ras GTPase-activating protein. *Mol Cell Biol* **11**(4)**:** 1804-1812

Morris DP, Greenleaf AL (2000) The splicing factor, Prp40, binds the phosphorylated carboxyl-terminal domain of RNA polymerase II. *J Biol Chem* **275**(51)**:** 39935-39943.

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**(4)**:** 536-540

Nilges M, Macias MJ, O'Donoghue SI, Oschkinat H (1997) Automated NOESY Interpretation with Ambiguous Distance Restraints: The Refined NMR Solution Structure of the Pleckstrin Homology Domain from β-Spectrin. *J Mol Biol* **269:** 408-422

Pascual J, Castresana J, Saraste M (1997) Evolution of the spectrin repeat. *Bioessays* **19**(9)**:** 811-817

Pearson JL, Robinson TJ, Munoz MJ, Kornblihtt AR, Garcia-Blanco MA (2008) Identification of the cellular targets of the transcription factor TCERG1 reveals a prevalent role in mRNA processing. *J Biol Chem* **283**(12)**:** 7949-7961

Phatnani HP, Greenleaf AL (2006) Phosphorylation and functions of the RNA polymerase II CTD. *Genes Dev* **20**(21)**:** 2922-2936

Qiao F, Song H, Kim CA, Sawaya MR, Hunter JB, Gingery M, Rebay I, Courey AJ, Bowie JU (2004) Derepression by depolymerization; structural insights into the regulation of Yan by Mae. *Cell* **118**(2)**:** 163-173

Ramirez-Espain X, Ruiz L, Martin-Malpartida P, Oschkinat H, Macias MJ (2007) Structural characterization of a new binding motif and a novel binding mode in group 2 WW domains. *J Mol Biol* **373**(5)**:** 1255-1268

Ridley AJ, Self AJ, Kasmi F, Paterson HF, Hall A, Marshall CJ, Ellis C (1993) rho family GTPase activating proteins p190, bcr and rhoGAP show distinct specificities in vitro and in vivo. *EMBO J* **12**(13)**:** 5151-5160

Roof RW, Haskell MD, Dukes BD, Sherman N, Kinter M, Parsons SJ (1998) Phosphotyrosine (p-Tyr)-dependent and -independent mechanisms of p190 RhoGAP-p120 RasGAP interaction: Tyr 1105 of p190, a substrate for c-Src, is the sole p-Tyr mediator of complex formation. *Mol Cell Biol* **18**(12)**:** 7052-7063

Roy AL (2007) Signal-induced functions of the transcription factor TFII-I. *Biochim Biophys Acta* **1769**(11-12)**:** 613-621

Sacristan C, Tussie-Luna MI, Logan SM, Roy AL (2004) Mechanism of Bruton's tyrosine kinase-mediated recruitment and regulation of TFII-I. *J Biol Chem* **279**(8)**:** 7147-7158

Sanchez-Alvarez M, Goldstrohm AC, Garcia-Blanco MA, Sune C (2006) Human transcription elongation factor CA150 localizes to splicing factor-rich nuclear speckles and assembles transcription and splicing components into complexes through its amino and carboxyl regions. *Mol Cell Biol* **26**(13)**:** 4998-5014

Saunders MW, A. Kirkwood, J.G. (1957) *J Am Chem Soc* **79**

Settleman J, Albright CF, Foster LC, Weinberg RA (1992) Association between GTPase activators for Rho and Ras families. *Nature* **359**(6391)**:** 153-154

Shen CH, Chen HY, Lin MS, Li FY, Chang CC, Kuo ML, Settleman J, Chen RH (2008) Breast tumor kinase phosphorylates p190RhoGAP to regulate rho and ras and promote breast carcinoma growth, migration, and invasion. *Cancer Res* **68**(19)**:** 7779-7787

Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* **11**(9)**:** 739-747

Smith MJ, Kulkarni S, Pawson T (2004) FF domains of CA150 bind transcription and splicing factors through multiple weak interactions. *Mol Cell Biol* **24:** 9274-9285

Spikes DA, Kramer J, Bingham PM, Van Doren K (1994) SWAP pre-mRNA splicing regulators are a novel, ancient protein family sharing a highly conserved sequence motif with the prp21 family of constitutive splicing proteins. *Nucleic Acids Res* **22**(21)**:** 4510-4519

Staunton D, Schlinkert R, Zanetti G, Colebrook SA, Campbell ID (2006) Cell-free expression and selective isotope labelling in protein NMR. *Magn Reson Chem* **44 Spec No:** S2-9

Sun X, Zhao J, Kylberg K, Soop T, Palka K, Sonnhammer E, Visa N, Alzhanova-Ericsson AT, Daneholt B (2004) Conspicuous accumulation of transcription elongation repressor hrp130/CA150 on the intron-rich Balbiani ring 3 gene. *Chromosoma* **113**(5)**:** 244-257

Sune C, Garcia-Blanco MA (1999) Transcriptional cofactor CA150 regulates RNA polymerase II elongation in a TATA-box-dependent manner. *Mol Cell Biol* **19**(7)**:** 4719-4728

Sune C, Hayashi T, Liu Y, Lane W, Young R, Garcia-Blanco MA (1997) CA150, a nuclear protein associated with the RNA polymerase II holoenzyme, is involved in Tat-activated human immunodeficiency virus type 1 transcription. *Mol Cell Biol* **17:** 6029-6039

Tcherkezian J, Lamarche-Vane N (2007) Current knowledge of the large RhoGAP family of proteins. *Biol Cell* **99**(2)**:** 67-86

Teigelkamp S, Newman AJ, Beggs JD (1995) Extensive interactions of PRP8 protein with the 5' and 3' splice sites during splicing suggest a role in stabilization of exon alignment by U5 snRNA. *EMBO J* **14**(11)**:** 2602-2612

Thompson JD, Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* **24:** 4876-4882

Vincent K, Wang Q, Jay S, Hobbs K, Rymond BC (2003) Genetic interactions with CLF1 identify additional pre-mRNA splicing factors and a link between activators of yeast vesicular transport and splicing. *Genetics* **164**(3)**:** 895-907

Wang Q, Hobbs K, Lynn B, Rymond BC (2003) The Clf1p splicing factor promotes spliceosome assembly through N-terminal tetratricopeptide repeat contacts. *J Biol Chem* **278**(10)**:** 7875-7883

Wellings DA, Atherton E (1997) Standard Fmoc protocols. *Methods Enzymol* **289:** 44-67

Wiesner S, Stier G, Sattler M, Macias MJ (2002) Solution structure and ligand recognition of the WW domain pair of the yeast splicing factor Prp40. *J Mol Biol* **324**(4)**:** 807-822

Zhang D, Childs G (1998) Human ZFM1 protein is a transcriptional repressor that interacts with the transcription activation domain of stage-specific activator protein. *J Biol Chem* **273**(12)**:** 6868-6877

Zhang D, Paley AJ, Childs G (1998) The transcriptional repressor ZFM1 interacts with and modulates the ability of EWS to activate transcription. *J Biol Chem* **273**(29)**:** 18086-18091

Zhou MM, Meadows RP, Logan TM, Yoon HS, Wade WS, Ravichandran KS, Burakoff SJ, Fesik SW (1995) Solution structure of the Shc SH2 domain complexed with a tyrosine-phosphorylated peptide from the T-cell receptor. *Proc Natl Acad Sci U S A* **92**(17)**:** 7784-7788

# APPENDIXS

# A1: Recipes for the production of $^{15}N$ / $^{13}C$ – labelled proteins

**M9 minimal medium (1 litre 10x stock solution):**

| | |
|---|---|
| $Na_2HPO_4$ | 60 g |
| $KH_2PO_4$ | 30 g |
| NaCl | 5 g |
| $NH_4Cl$ / $^{15}NH_4Cl$ | 5 g (for unlabelled / $^{15}N$-labelled medium) |

Add $H_2O$ up to 1 litre. Sterilize by autoclave

**M9 minimal medium (1 litre)**

| | |
|---|---|
| M9 medium 10X | 100 ml |
| Trace elements 100X | 10 ml |
| Glucose 20 % | 20 ml |
| $MgSO_4$ 1M | 1 ml |
| $CaCl_2$ 1M | 300 µl |
| Biotin 1mg/ml | 1 ml |
| Thiamin 1mg/ml | 1 ml |
| Kanamycin 50 mg/ml * | 1 ml |

* (or another suitable antibiotic)

Add $H_2O$ up to 1 litre. Sterilize by filtration

**$^{15}N$- M9 minimal medium (1 litre)**

| | |
|---|---|
| **$^{15}N$**-M9 medium 10X | 100 ml |
| Trace elements 100X | 10 ml |
| Glucose 20 % | 20 ml |
| $MgSO_4$ 1M | 1 ml |
| $CaCl_2$ 1M | 300 µl |
| Biotin 1mg/ml | 1 ml |
| Thiamin 1mg/ml | 1 ml |
| Kanamycin 50 mg/ml * | 1 ml |

* (or another suitable antibiotic)

Add $H_2O$ up to 1 litre. Sterilize by filtration

**$^{15}N$- $^{13}C$- M9 minimal medium (1 litre)**

| | |
|---|---|
| **$^{15}N$**-M9 medium 10X | 100 ml |
| Trace elements 100X | 10 ml |
| **$^{13}C$**- glucose | 2 g |
| $MgSO_4$ 1M | 1 ml |
| $CaCl_2$ 1M | 300 µl |
| Biotin 1mg/ml | 1 ml |
| Thiamin 1mg/ml | 1 ml |
| Kanamycin 50 mg/ml * | 1 ml |

* (or another suitable antibiotic)

Add $H_2O$ up to 1 litre. Sterilize by filtration

**Trace elements (1 litre 100X stock solution)**

| | |
|---|---|
| EDTA | 5 g |
| $FeCl_3$ (6 $H_2O$) | 0.83 g |
| $ZnCl_2$ | 84 mg |
| $CuCl_2$ (2 $H_2O$) | 13 mg |
| $CoCl_2$ | 10 mg |
| $H_3BO_3$ | 10 mg |
| $MnCl_2$ | 1.6 mg |

First dissolve the EDTA in 800 ml water and adjust the pH to 7.5. Then add other components and add $H_2O$ up to 1 litre. Sterilize by filtration.

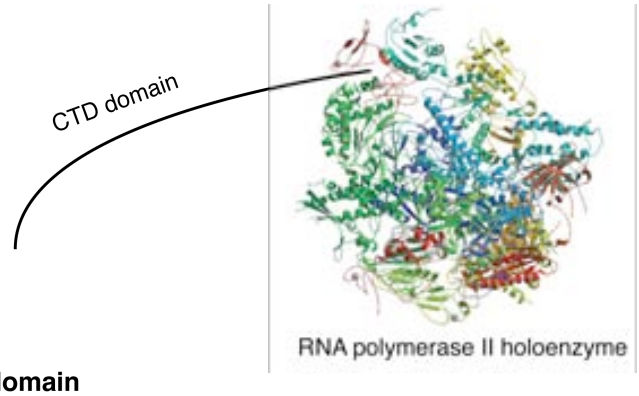
## A2: The C-terminal domain (CTD) of RNA-polymerase II

(reviewed in (Meinhart et al, 2005; Phatnani & Greenleaf, 2006))

The C-terminal domain (CTD) of the RNA polymerase II is a tail-like extension attached to the holoenzyme of the polymerase. CTD plays an indispensable role for life, as cells lacking at least two thirds of the repeats are unviable.

The free CTD is flexible and mainly unstructured and is composed of repeated motifs of the consensus sequence YSPTSPS that range from 26 (in yeast) to 52 (in human). (Fig. 40)

The CTD repeats are subjected to phosphorylation, which occurs principally at Ser 2 and Ser 5.




**CTD domain**

YSPTSPS**YSPTSPS**YSPTSPS

**phospho-CTD domain**

YpSPTpSPS**YpSPTpSPS**YpSPTpSPS

**Figure 40: Picture of the RNA-polymerase II holenzyme and its C-terminal domain.** Below, the sequence of the CTD and the phospho-CTD repeats are indicated, with the consensus motif underlined.

The CTD repeats bind numerous nuclear factors and so, serve as a platform to link nuclear events to transcription. Moreover, the association of a given factor to the CTD depends on its phosphorylation pattern, which changes during the movement of RNA-pol II from one gene to the next. For example, the coupling of the pre-mRNA processing to transcription requires the recruitment of several factors to the CTD. Some of these CTD associated factors include the Cgt1 domain of the RNA capping enzyme, which interacts with the CTD when the repeats are phosphorylated on Ser 5, or the poly(A)-dependent 3'-RNA processing factor Pcf11 which needs the CTD to be phosphorylated on Ser 2.

Another binder is the prolyl isomerase Pin1, which associates with the phospho-CTD via its WW domain. Conversely, the first WW of the ubiquitin ligase Rsp5 has a greater affinity for the CTD compared to that for the phospho-CTD.

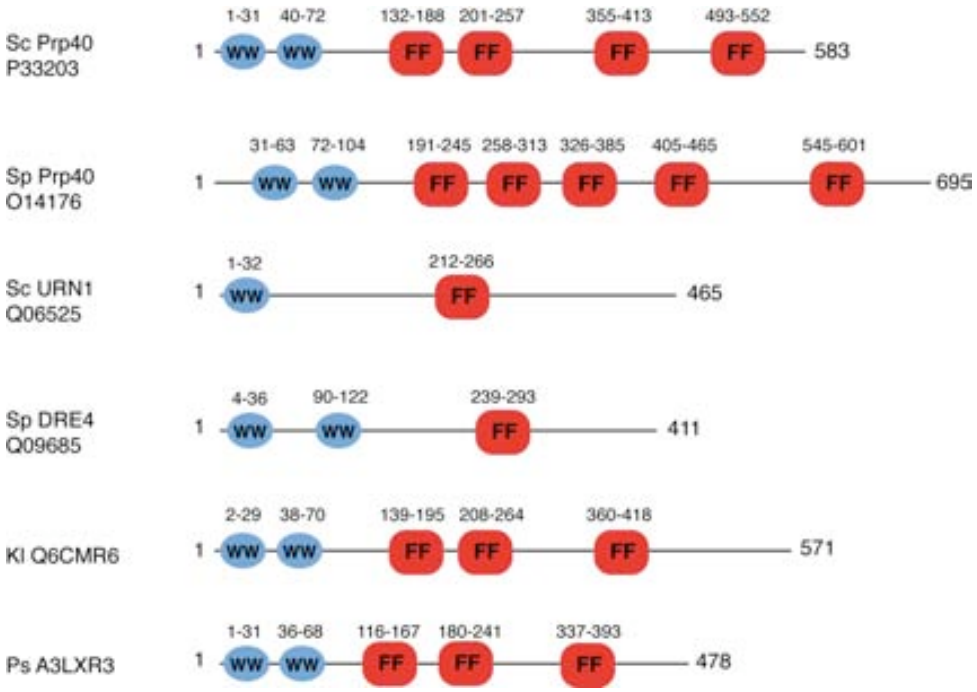
Despite the information available and the recent founds, there are still many open questions regarding aspects like the binding mode of the associated factors to the CTD or the coordination of kinases and phosphatases to produce changes in the phosphorylation pattern.

## A3: List of entries from the main FF-containing protein families

Sequence numbering and domain composition of the entries are shown according to SMART database descriptions (http://smart.embl.de/). SwissProt /UniProt (http://www.uniprot.org/) accession numbers are indicated for the entries. The domains present in the different entries are: FF- *FF domain,* WW- *WW domain,* RhoGAP- *Rho GTPase Activating Protein domain,* small GTPase- *small GTPase of the Ras superfamily domain*

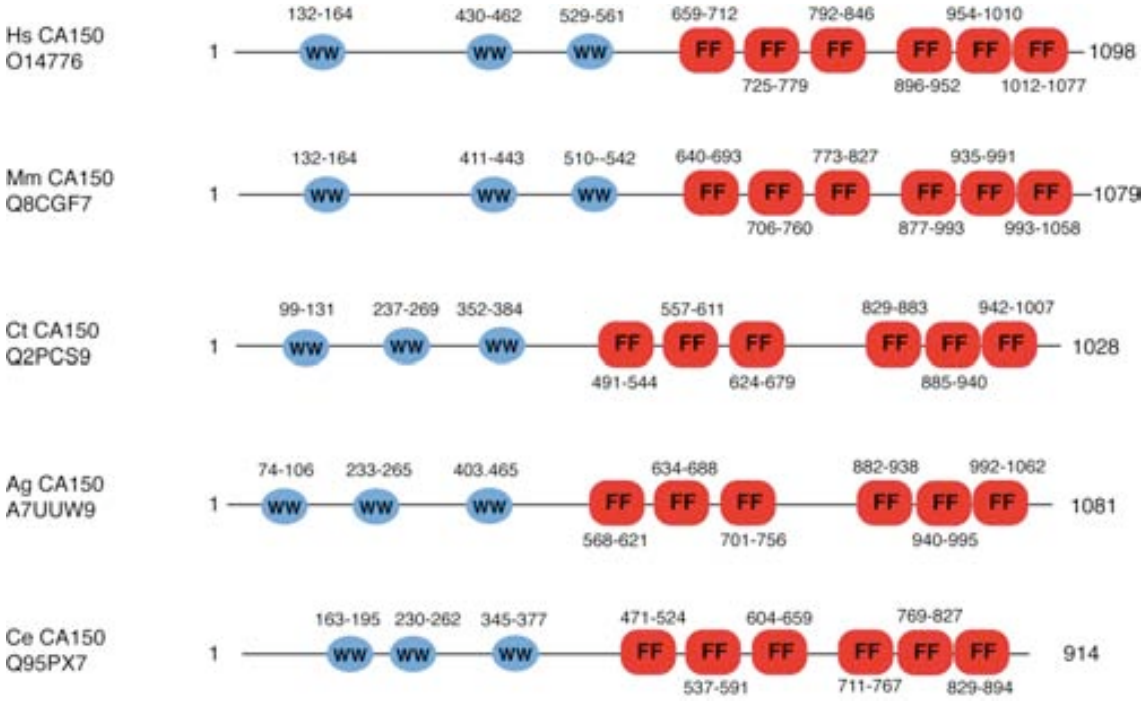
### Yeast FF-containing proteins:

P33203*, Saccharomyces cerevisiae* Prp40; O14176, *Saccharomyces pombe* Prp40; Q06525, *Saccharomyces cerevisiae* URN1; Q09685, *Saccharomyces pombe* DRE4; Q6CMR6, *Kluyveromyces lactis* KLLA0E18239p; A3LXR3, *Pichia stipitis* Pre-mRNA processing protein



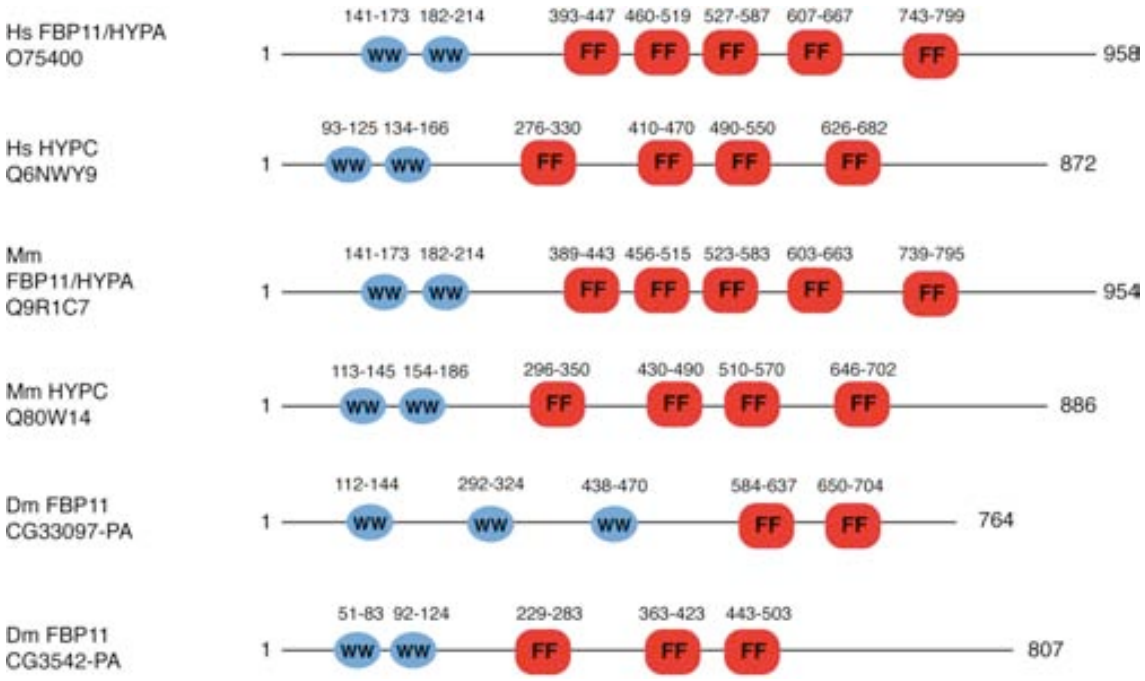
## CA150 protein family:

O14776, *Homo sapiens* CA150; Q8CGF7, *Mus musculus* CA150; Q2PCS9 *Chironomus tentans* hrp130; A7UUW9 *Anopheles gambiae* CA150; Q95PX7 *Caenorhabditis elegans* CA150



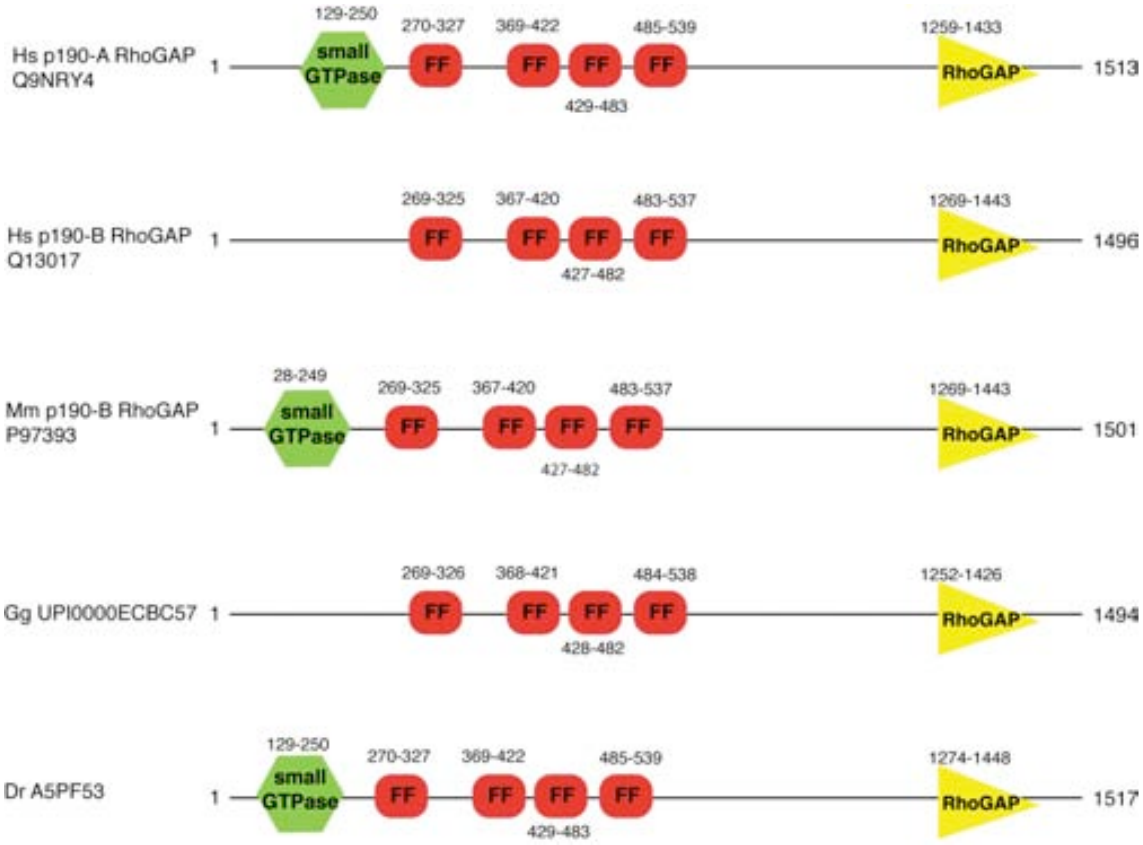

## FBP11 / HYPC protein families:

O75400, *Homo sapiens* FBP11; Q6NWY9, *Homo sapiens* HYPC; Q9R1C7, *Mus musculus* FBP11; CG33097-PA, *Drosophila melanogaster* FBP11; CG3542-PA, *Drosophila melanogaster* FBP11



## p190 Rho-GAP protein family:

Q9NRY4, *Homo sapiens* p190-A RhoGAP; Q13017 *Homo sapiens* p190-B RhoGAP; P97393 *Mus musculus* p190-B RhoGAP; UPI0000ECBC57, *Gallus gallus* similar to GRLF1; A5PF53 *Danio rerio* similar to GRLF1



## A4: List of constructs used during this thesis

All the domains used in this thesis contain additional N-terminal residues resulting from the purification His tag (MKHHHHHHPMG) or the TEV cleavage of the fusion protein (GAM or GAMG). Sequence numbering according to the entire protein, molecular weight and overall pKas of the constructs are indicated. All yeast FF clones except that of Prp40FF3 were provided by Dr. A.Gasch.

### Yeast Prp40 / URN1 FF domains:

**URN1 FF** (212-266):
GAM DIDERNIFFE LFDRYKLDKF STWSLQSKKI ENDPDFYKIR DDTVRESLFE EWCGE
Mw: 7117.8 Da / overall $pK_a$: 4.5

**Prp40 FF1** (134-187)
GAM EAEKEFITML KENQVDSTWS FSRIISELGT RDPRYWMVDD DPLWKKEMFE KYLSNR
Mw: 7150.1 Da / overall $pK_a$: 4.8

**Prp40 FF2** (200-260):
GAM ETSKFKEAFQ KMLQNNSHIK YYTRWPTAKR LIADEPIYKH SVVNEKTKRQ TFQ DYIDTLID
Mw: 7667.7 Da / overall $pK_a$: 9.3

**Prp40 FF1-2** (134-260):
MKHHHHHHPM EAEKEFITML KENQVDSTWS FSRIISELGT RDPRYWMVDD DPLWK KEMFE KYLSNRSADQ LLKEHNETSK FKEAFQKMLQ NNSHIKYYTR WPTAK RLIAD EPIYK HSVVNEKTKR QTFQDYIDT LID
Mw: 16727.9 Da / overall $pK_a$: 7.1

**Prp40 FF3** (346-412):
GAM ELRLRNYTRD RIARDNFKSL LREVPIKIKA NTRWSDIYPH IKSDPRFLHM LGRNG SSCLD LFLDFVD
Mw: 10583.2 Da / overall $pK_a$: 9.9

**Prp40 FF4** (488-552):

GAM NERRILEQKK HYFWLLLQRT YTKTGKPKPS TWDLASKELG ESLEYKALGD EDN IRRQIFE DFKPE

Mw: 8145.2 Da / overall $pK_a$: 8.2

## CA150 FF domains:

(residues in red indicate mutations from the original sequence. We did not revert them because in all cases were conservative changes according to the FF alignments. In the case of FF4 construct we had an insertion of several residues after the C-terminal of the domain that did not affect domain folding)

**CA150 FF1** (651-715):

GAM ARERAIVPLE ARMKQFKDML LERGVSAFST WEKELHKIVS DPRYLLLNPK ERK QVFDQYV KTRAE

Mw: 8110.5 Da / overall $pK_a$: 9.7

**CA150 FF2** (720-782):

GAM EKKNKIMQAK EDFKKMMEEA KFNPRATFSE FAAKHAKDSR FKAIEKMKDR EA LFNEFVAA ARK

Mw: 7701.9 Da / overall $pK_a$: 9.8

**CA150 FF1-2** (651-782):

GAM ARERAIVPLE ARMKQFKDML LERGVSAFST WEKELHKIVF DPRYLLLNPK ERK QVFDQYV KTRAEEERRE KKNKIMQAKE DFKKMMEEAK FNPRATFSE AAKHAKDSRF KAIEKMKDRE ALFNEFVAAA RK

Mw: 16105.7 Da / overall $pK_a$: 9.8

**CA150 FF3** (784-853):

GAM EKEDSKTGGE KIKSDFFELL SNHHLDSQSR WSKVKDKVES DPRYKAVDSS SMR EDLFKQY IEKIAKNLDS

Mw: 8439.3 Da / overall $pK_a$: 6.1

**CA150 FF4** (890-956):

MKHHHHHHPM EREQHKREEA IQNFKALLSD MVRSSDVSWS DTRRTLRKDH RWESG SLLER EEKEKLFNEH IEALTKK RESTLGNFWMKLLQLP

Mw: 9449.6 Da / overall $pK_a$: 8.2

**CA150 FF5** (951-1009):
MKHHHHHHPM EALTKKKREH FRQLLDETSA ITLTSTWKEV KKIIKEDPRC IKFSSSDR KK QREFEEYIR
Mw: 8510.8 Da / overall $pK_a$: 9.7

**CA150 FF6** (1010-1077):
MKHHHHHHPM DKYITAKADF RTLLKETKFI TYRSKKLIQE SDQHLKDVEK ILQNDKR YLV LDCVPEERRK LIVAYVDD
Mw: 9499.8 Da / overall $pK_a$: 9.0

## p190-A RhoGAP FF domains:

**p190-A RhoGAP FF1** (267-327):
GAMG SQQIATAKDK YEWLVSRIVK NHNENWLSVS RKMQASPEYQ DYVYLEGTQK AKKLFLQHIH R
Mw: 7650.7 Da / overall $pK_a$: 9.6

**p190-A RhoGAP FF1 extended** (267-331):
GAMG SQQIATAKDK YEWLVSRIVK NHNENWLSVS RKMQASPEYQ DYVYLEGTQK AKKLFLQHIH RLKHE
Mw: 8158.3 Da / overall $pK_a$: 9.5

**p190-A RhoGAP FF4** (485-542):
GAMG IDKAKEEFQE LLLEYSELFY ELELDAKPSK EKMGVIQDVL GEEQRFKALQ K LQAERD
Mw: 7077.0 Da / overall $pK_a$: 4.6

**N-terminal TFII-I** (1-89):
GA MAQVAMSTLP VEDEESSESR MVVTFLMSAL ESMCKELAKS KAEVACIAVY ETDV FVVGTE RGRAFVNTRK DFQKDFVKYC VEEEEKAAE
Mw: 10126.5 Da / overall $pK_a$: 4.6

**N-terminal TFII-I** (1-60):
GA MAQVAMSTLP VEDEESSESR MVVTFLMSAL ESMCKELAKS KAEVACIAVY ETDV FVVGTE
Mw: 6650.6 Da / overall $pK_a$: 4.3

**N-terminal TFII-I** (21-89):

GA MVVTFLMSAL ESMCKELAKS KAEVACIAVY ETDVFVVGTE RGRAFVNTRK DFQKDFVKYC VEEEEKAAE

Mw: 7948.1 Da / overall $pK_a$: 4.9

**N-terminal TFII-I** (21-60):

GA MVVTFLMSAL ESMCKELAKS KAEVACIAVY ETDVFVVGTE

Mw: 4472.2 Da / overall $pK_a$: 4.6

# A5: Publications related to this thesis

## Chapter 1

Bonet, R., Ramirez-Espain, X. and Macias, M.J. *Solution structure of the URN1 splicing factor FF domain: Comparative analysis of charge distributions in FF domain structures; FFs and SURPs, two domains with a similar fold.* Proteins. 2008 Dec; 73(4): 1001-9

Bonet, R., Ruiz, L., Morales, B., Macias, M.J. *Solution structure of the fourth FF domain of yeast Prp40 splicing factor* submitted to Proteins

## Chapter 3

Bonet R, Ruiz L, Aragón E, Martín-Malpartida P, Macias MJ. *NMR structural studies on human p190-A RhoGAP FF1 revealed that domain phosphorylation by the PDGF-receptor alpha requires its previous unfolding.* J Mol Biol. 2009 Jun; 389 (2): 230-237