



Universitat Autònoma de Barcelona

ADVERTIMENT. L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons:  http://cat.creativecommons.org/?page_id=184

ADVERTENCIA. El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons:  <http://es.creativecommons.org/blog/licencias/>

WARNING. The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  <https://creativecommons.org/licenses/?lang=en>



Universitat Autònoma
de Barcelona



**Auxin Response Factors:
integrating auxin signalling with DNA
recognition and epigenetics**

Doctoral thesis presented by

Isidro Crespo García

Graduated in Biochemistry

To apply for the degree of Doctor in Biochemistry, Molecular
Biology and Biomedicine by Universitat Autònoma de Barcelona.

Thesis performed at the ALBA Synchrotron Light Source, co-
directed by **Dr. Roeland Boer** and **Prof. Dolf Weijers**.

A handwritten signature in blue ink, consisting of a large, stylized 'R' followed by a long horizontal stroke.

Dr. Roeland Boer

A handwritten signature in blue ink, featuring a stylized 'D' and 'W' with a long horizontal stroke.

Prof. Dolf Weijers

Isidro Crespo García

November 15th, 2020

Agraïments

En arribar al final d'aquesta etapa no puc evitar mirar enrere i observar el camí recorregut. Aquest camí que ara em porta a presentar aquest treball ha sigut dur en alguns moments, però ha estat ple de gent maca i bones experiències que m'han ajudat a continuar. Aquesta feina no hauria sigut possible sense el suport incondicional dels meus directors. Dolf, thank you very much for the opportunity of being part of this research, for me was an amazing experience to work with you. At the beginning I was a bit afraid of working with a plant system, something new for me at that moment. I really enjoyed this project and I hope to continue working with you. I a en Roeland, moltes gràcies per la teva comprensió, fins i tot en els moments difícils en els que hauria escampat la boira. La teva paciència i els teus consells han millorat la qualitat d'aquesta feina. El teu rigor científic és un exemple a seguir. Per descomptat, també agraeixo el tracte VIP a la línia, omplint foradets dels usuaris per ficar els meus cristalls. I no cal dir res de les jornades maratonianes de correccions...

Als meus companys de laboratori i d'oficina, res no és igual sense vosaltres, el laboratori està ben avorrit... Recordo aquelles partides de ping-pong el divendres abans de torna a casa amb en Harold "cachoperro", el proctodoctor Albert, la Nerea, l'Anna Cuppari, l'Annia, la Laia "funcionària"... Sense oblidar la Marta, que no cal ser predoc per poder gaudir d'alguna partideta de tant en tant. He tingut la sort de compartir aquest temps amb molta més gent d'ALBA, que m'han fet veure les coses d'una altra manera. A les becàries Alba, Mari Carmen, Carla, Mònica i Anabel... Haver sigut el vostre tutor de pràctiques ha sigut una gran experiència personal. M'heu ensenyat com explicar els conceptes, a vegades complexos del laboratori, per poder transmetre-ho millor. Espero que tingueu molta sort en el vostre futur. A la Judith, vas ser-hi en el moment en el que necessitava veure les coses d'una altra manera, i t'estic molt agraït per això. També a en Damià, per recordar-me que no es pot oblidar res per bàsic que sigui (ARG!). Als companys de la línia Xavi, Fernando i Bàrbara, m'ha agradat molt participar amb vosaltres en experiments bojós a la línia i espero haver-vos ajudat quan m'heu necessitat. Als "flores", per la bona conversa que tenen i l'ajuda que donen en tots els àmbits científics. A la Daimí i a l'Inma, no sé què hauria fet sense vosaltres. Sóc un desastre amb la burocràcia d'ALBA, però amb vosaltres la paperassa es fa més lleugera. Part d'aquesta feina també l'agraeixo a la meva comissió de seguiment, que amb molt de criteri sempre m'han aconsellat bé per on seguir, en especial a en Manel, que tot i jubilar-se va accedir a continuar presidint la comissió.

Per suposat, haver continuat en ciència es en gran part dels ADHs, que em van mostrar lo maco d'aquesta feina. A en Jaume i en Xavier, per haver-me acceptat al seu grup i haver confiat sempre en mi. A en Sergio, els seus "podries" em van obrir la ment i descobrir que mai es pot treure una conclusió sense fer tots els experiments possibles. A la Raquel, que vam començar les pràctiques el mateix dia, encara recordo quan es va perdre una cubeta i anàvem bojós buscant-la. Compartir el laboratori amb tu va ser una gran experiència, la teva forma de veure la vida i els teus consells van ser essencials. Ets una gran professional i una gran amiga. A en Joan i en Iago, estic molt content d'haver sigut el vostre "lacayo", em va iniciar en aquest món. I per descomptat en Julio, quina sort haver sigut company teu. No només em vas transmetre el teu rigor si no que em vas ajudar en el pitjor moment. Us estaré sempre agraït a tu i la Mayra.

Per últim, vull dedicar aquest treball a la meva família. A tu "yaya", si vaig començar en això va ser per tu. De petit em deies que com em cabien tantes lletres al cap. Espero que allà on siguis estiguis contenta de com ens van les coses. Als tiets i les cosines, desitjo tornar a la normalitat per trobar-nos de nou. A la meva mare, que tot i que pateixes massa per mi i que t'interessen altres línies de recerca més mèdiques, sempre em dones suport en tot el que faig i puc comptar amb la

teva ajuda. Sense tu no ho hauria aconseguit. A en Sergio i en Jerry, se us troba a faltar encara que no estiguem gaire lluny. Els anys que vam viure “independents” han sigut els millors, i heu sigut una peça fonamental per poder tirar endavant aquest projecte. Sóc molt afortunat tenint un germà i un gat així. Als peques, no tan peques, no us rendiu mai, s’aconsegueixen més cosses per tossut que no pas per brillant. I que sapiguen que sempre em teniu disponible pel que sigui. A l’Espí, un amic gairebé de la família. Hem de recuperar la bici els diumenges i aconseguir fer d’una vegada la calçotada. I a l’Anna, la teva comprensió i suport són un pilar fonamental. Escoltes les meves idees quan em poso creatiu, encara que la biologia no sigui el teu camp. M’aconselles i m’ajudes encara que no t’ho demani, i sempre estàs disposada a fer un tomb quan la situació ho requereix. Amb tu he après a estimar la natura, i no cal dir que això es una part fonamental d’aquest treball. Gràcies a tu aquesta feina s’ha fet molt més senzilla. M’alegro molt formis part de la meva vida.

Sóc molt afortunat.

A la meva família

Abstract

Auxins are the main phytohormones governing plant growth and development. From this group of hormones, indole-3-acetic acid is the main physiological auxin regulating plant growth. The main regulatory path of gene regulation by auxin is the **Nuclear Auxin Pathway**, NAP. The final effectors of this signalling pathway are the DNA binding transcription factors **Auxin Response Factors** (ARFs). ARFs has been linked to auxin and gene expression for decades, but their structures have only been recently obtained. The crystallographic structures of the DNA Binding Domain (ARF-DBD) of two divergent *Arabidopsis thaliana* ARFs revealed an exceptionally high conservation of the DNA contact points, so specificity of gene expression was unlikely to reside in the DBD. Those structures suggested that specificity for gene selection may require dimerization and proper spacing of DNA binding elements for ARF binding, a hypothesis known as the “**molecular calliper**” model. We present in **chapter 2** new structures of the *Marchantia polymorpha* ARF2 in complex of DNA sequences. Our results show that, despite *Marchantiophyta* and *Brassicaceae* diverged more than 400 million years ago, the structure is conserved, even in the DNA contacting points. We also show that the relative position of the DBD subdomains may determine spacing selectivity, which are different for AtARF1 compared to AtARF5 and MpARF2. AtARF1 B3 domains are more distant, showing a preference for higher AuxRE spacing. The only difference that we have found in the structures to explain this effect is located in $\alpha 1$, which is longer in AtARF1 and establishes more contacts that lock the relative position of the subdomains compared to the other ARFs. The structural predictions based on intersubdomain relative position were confirmed *in vitro* in **chapter 3**. In this chapter we also found that Apo ARFs are able to dimerize *in vitro*, but the main driving force for dimerization is DNA interaction. The dimerization interface is also very conserved, and it was found that ApoARFs can heterodimerize, even in distantly related ARFs. All these elements agree with a gene regulation governed by the molecular calliper model. We also believe that, although the molecular calliper can partially explain ARF specificity, it is not enough to explain all the genetic variability triggered by auxins in complex organisms. We propose that the molecular calliper may function as an ancestral mechanism of specificity selection. However, other specificity selectors should exist to explain the regulation of ARF-controlled genes. Our structures also propose an interesting explanation for high DNA affinity, where His136 (AtARF1 numbering) sidechain flip allows a high affinity interaction with certain DNA sequences. This histidine is not present in Class C ARFs, representing a difference in DNA affinity between classes A/B and C.

Our structural analysis is also expanded to the **Ancillary Domain** (AD) of ARFs, a DBD subdomain with no previous function assigned. The fold of the Ancillary Domain is related to Tudor domains, a member of the Royal Family (RF) of **histone methyllysine and methylarginine readers**. Ancillary Domains shared elements with multiple RF subfamilies, which prompted us to compare the AD with known RF proteins. In **chapter 4** we review the Royal Family classification, summarizing the elements required for a protein to be part of this superfamily. The work presented in **chapter 4** allowed us to analyse whether ARFs comply with the RF family characteristics. We demonstrate in **chapter 5** that ARF-AD retains all the structural elements to be a functional Histone posttranslational modification reader, and that this domain is not sequence related with any other protein not being ARFs or plant proteins. We have seen that the binding cage in ARFs is always covered by a basic amino acid, which would prevent the binding of molecules to the ARF-AD. Furthermore, we found that the “Taco fold” of the Dimerization Domain resembles to the fold of many members of the RF superfamily, with a non-functional HC. In other RF domains, as in extended Tudor and extended chromodomains, the two RF-like modules provide different functionalities. In ARFs, the DD adds dimerization capability to the AD. A clear classification of Ancillary domains into the existing RF protein families was not possible as Ancillary Domains

shared elements with multiple RF subfamilies. In consequence, we propose that ARFs constitute a new RF family specific of plants and coined the term “**Steward domain**” to refer to the combined RF-like domains found in the DD and AD of the ARFs. In **chapter 6**, we demonstrate that AtARF1-DBD interacts with several histone peptides, and that the pH affects the interaction. In addition, array analyses provides a landscape of interactions of AtARF1-DBD with modified Histone-derived peptides, which suggests an *in vivo* histone reading module functionality, that would select genes under auxin presence based on the epigenetic status of the nucleosome. This may represent the missing piece in the ARF regulation puzzle.

Finally, in **chapter 7** we discuss and summarize all the information gathered in this work and future directions and considerations are presented.

Resum

Les auxines són les principals fitohormones que regulen el creixement i desenvolupament vegetal. D'aquest grup d'hormones, l'àcid indol-3-acètic és la principal auxina fisiològica. La principal via reguladora de la regulació gènica per auxina és la **Ruta Nuclear de l'Auxina**, NAP. Els efectors finals d'aquesta via de senyalització són uns factors de transcripció coneguts com **Factors de Resposta a les Auxines** (ARF). Les ARF s'han relacionat amb les auxines i l'expressió de gens durant dècades, però les seves estructures van ser resoltes recentment. Les estructures cristal·logràfiques del domini d'unió a l'ADN (ARF-DBD) de dues ARF divergents d'*Arabidopsis thaliana* van revelar una conservació excepcionalment alta dels punts de contacte de l'ADN, de manera que era poc probable que l'especificitat de l'expressió gènica residís al DBD. Aquestes estructures van suggerir que l'especificitat per a la selecció de gens podria requerir de la dimerització i l'espaiat adequat dels elements d'unió a l'ADN per a la unió efectiva de les ARF, una hipòtesi coneguda com a model de "calibres moleculars". Presentem en el **capítol 2** noves estructures de *Marchantia polymorpha* ARF2 en complex amb ADN. Els nostres resultats mostren que, tot i que les *Marchantiophyta* i les *Brassicaceae* van divergir fa més de 400 milions d'anys, l'estructura es conserva, fins i tot en els punts de contacte amb l'ADN. També mostrem que la posició relativa dels subdominis del DBD pot determinar la selectivitat de l'espaiat, que són diferents per a AtARF1 en comparació amb AtARF5 i MpARF2. Els dominis B3 d'AtARF1 són més distants, mostrant una preferència per un major espaiat d'AuxRE. L'única diferència que hem trobat en les estructures per explicar aquest efecte es troba en $\alpha 1$, que és més llarg en AtARF1 i estableix més contactes que bloquegen la posició relativa dels subdominis en comparació amb els altres ARF. Les prediccions estructurals basades en la posició relativa entre subdominis es van confirmar *in vitro* al **capítol 3**. En aquest capítol també trobem que els Apo ARF són capaços de dimeritzar *in vitro*, però la principal impulsora de la dimerització és la interacció amb l'ADN. La interfície de dimerització també està molt conservada, ja que vam trobar que els ApoARF poden heterodimeritzar, fins i tot amb ARF distants en l'evolució. Tots aquests elements concorden amb una regulació gènica regida pel model del calibre molecular. També creiem que, encara que el calibre molecular pot explicar parcialment l'especificitat de l'ARF, no és suficient per explicar tota la variabilitat genètica desencadenada per les auxines en organismes complexos. Proposem que el calibre molecular pot funcionar com un mecanisme ancestral de selecció d'especificitat, tot i que no obstant això, hi hauria d'haver altres selectors d'especificitat per explicar la regulació de gens controlats per ARF. Les nostres estructures també proposen una explicació interessant per a l'alta afinitat de certes seqüències d'ADN, on el moviment de la cadena lateral His136 (numeració AtARF1) permet una interacció d'alta afinitat amb certes seqüències d'ADN. Aquesta histidina no està present en les ARF de classe C, el que representa una diferència en l'afinitat de l'ADN entre les classes A/B i C.

La nostra anàlisi estructural també s'amplia al domini auxiliar d'ARF (**Ancillary domain**, AD), un subdomini del DBD sense funció prèvia assignada. El plegament del domini auxiliar està relacionat amb els dominis Tudor, un membre de la "**Royal Family**" (RF) de lectors d'histones metilades a lisina i arginina. Els dominis auxiliars comparteixen elements estructurals amb múltiples subfamílies de RF, fet que ens va portar a comparar l'AD amb proteïnes de RF conegudes. Al **capítol 4** revisem la classificació de la RF, resumint els elements necessaris perquè una proteïna sigui part d'aquesta superfamília. El treball presentat al capítol 4 va permetre analitzar si les ARF compleixen amb les característiques de les RF. Vam demostrar al **capítol 5** que l'ARF-AD conté tots els elements estructurals per ser un lector funcional de modificacions postraduccionals d'histones, i que aquest domini no està relacionat en seqüència amb cap altra proteïna que no sigui ARF o proteïnes vegetals. Hem vist que el lloc d'unió als ARF-AD sempre està cobert per un aminoàcid bàsic, el que evitaria la unió de molècules a l'ARF-AD. A més, trobem que el plegament

del domini de dimerització (DD) de l'ARF-DBD s'assembla al plegament de molts membres de la superfamília RF, amb un lloc d'unió no funcional. En altres dominis de RF, com en els Tudor i cromodomini estesos, els dos mòduls similars a RF proporcionen diferents funcionalitats. Llavors en els ARF, el DD afegeixiria capacitat de dimerització a l'AD. No va ser possible una classificació clara dels dominis auxiliars d'ARF en les famílies de proteïnes de RF existents, ja que els dominis auxiliars compartien elements amb múltiples subfamílies de RF. En conseqüència, proposem que els ARF constitueixen una nova família de RF i encunyem el terme "domini Steward" per referir-nos als dominis combinats similars a RF que es troben en el DD i AD dels ARF. En el **capítol 6**, vam demostrar que AtARF1-DBD interactua amb diversos pèptids d'histones i que el pH afecta la interacció. A més, els anàlisis d'arrays ens van mostrar un gran ventall d'interaccions de AtARF1-DBD amb pèptids derivats d'histones modificades, el que suggereix una funcionalitat de lectura d'histones *in vivo*, que seleccionaria gens en presència d'auxines en funció de l'estat epigenètic del nucleosoma . Això pot representar la peça que falta en el trencaclosques de la regulació sota les ARF.

Finalment, en el **capítol 7** discutim i resumim tota la informació recopilada en aquest treball i es presenten les adreces i consideracions futures.

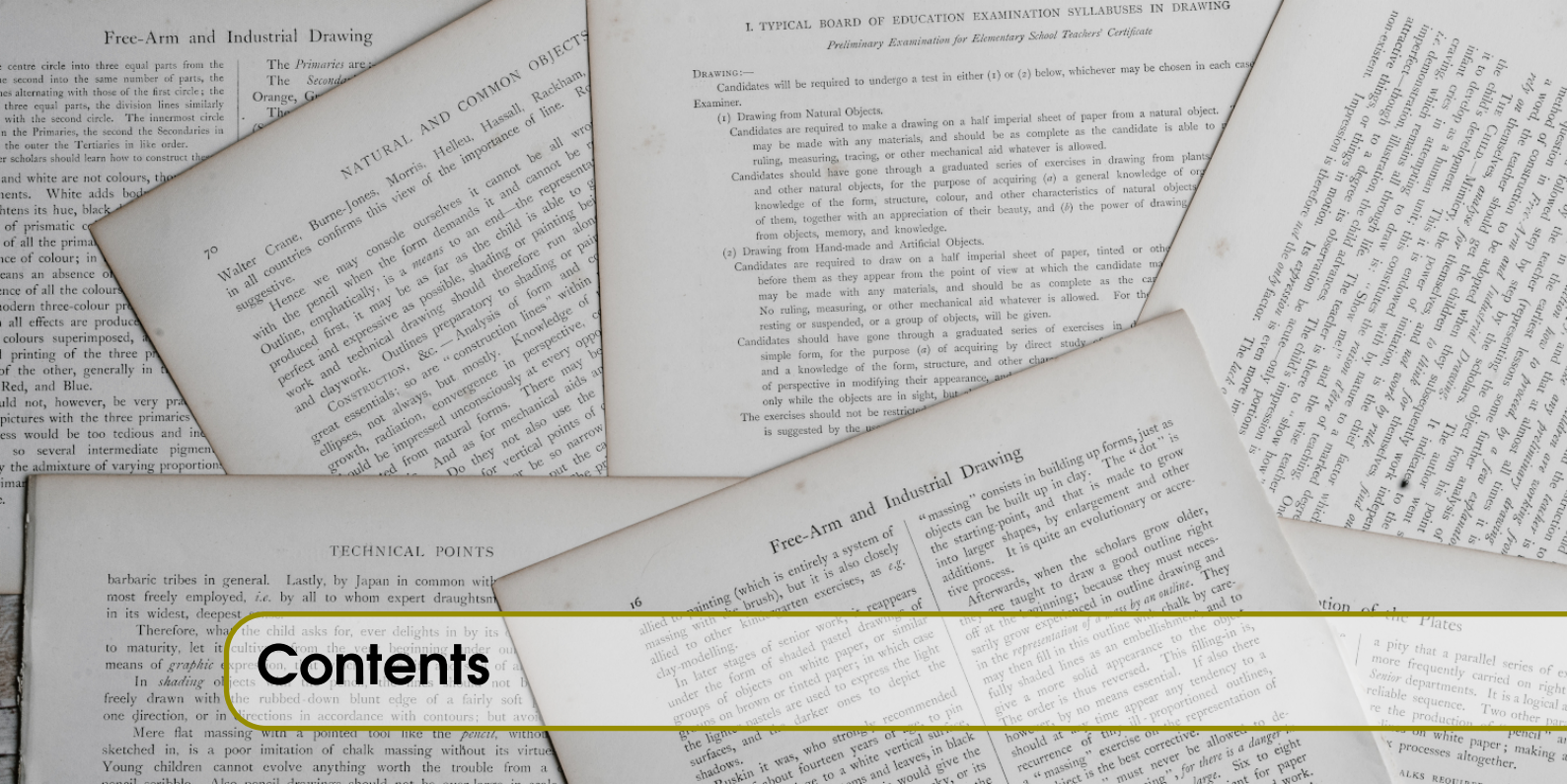
Resumen

Las auxinas son las principales fitohormonas que gobiernan el crecimiento y desarrollo vegetal. De este grupo de hormonas, el ácido indol-3-acético es la principal auxina fisiológica que regula el crecimiento de las plantas. La principal vía de regulación génica por auxina es la **ruta nuclear de la auxina**, NAP. Los efectores finales de esta vía de señalización son los factores de transcripción conocidos como **Factores de Respuesta a las Auxinas** (ARF). Los ARF se han relacionado con las auxinas y la expresión de genes durante décadas, pero sus estructuras se han podido resolver recientemente. Las estructuras cristalográficas del dominio de unión al ADN (ARF-DBD) de dos ARF divergentes de *Arabidopsis thaliana* revelaron una conservación excepcionalmente alta de los puntos de contacto del ADN, por lo que era poco probable que la especificidad de la expresión génica residiera en el DBD. Esas estructuras sugirieron que la especificidad para la selección de genes puede requerir la dimerización y el espaciado adecuado de los elementos de unión al ADN para la unión de ARF, una hipótesis conocida como modelo de "**calibres moleculares**". Presentamos en el **capítulo 2** nuevas estructuras de *Marchantia polymorpha* ARF2 en complejo con ADN. Nuestros resultados muestran que, a pesar de que las *Marchantiophyta* y las *Brassicaceae* divergieron hace más de 400 millones de años, la estructura se conserva, incluso en los puntos de contacto del ADN. También mostramos que la posición relativa de los subdominios DBD puede determinar la selectividad de espaciado, que son diferentes para AtARF1 en comparación con AtARF5 y MpARF2. Los dominios B3 de AtARF1 están más distantes, mostrando una preferencia por un mayor espaciado AuxRE. La única diferencia que hemos encontrado en las estructuras para explicar este efecto se encuentra en $\alpha 1$, que es más largo en AtARF1 y establece más contactos que bloquean la posición relativa de los subdominios en comparación con los otros ARF. Las predicciones estructurales basadas en la posición relativa entre subdominios se confirmaron *in vitro* en el **capítulo 3**. En este capítulo también encontramos que los Apo ARF son capaces de dimerizar *in vitro*, pero la principal fuerza impulsora de la dimerización es la interacción del ADN. La interfaz de dimerización también está muy conservada, y se encontró que los ApoARF pueden heterodimerizar, incluso en ARF evolutivamente distantes. Todos estos elementos concuerdan con una regulación génica regida por el modelo del calibre molecular. También creemos que, aunque el calibre molecular puede explicar parcialmente la especificidad del ARF, no es suficiente para explicar toda la variabilidad genética desencadenada por las auxinas en organismos complejos. Proponemos que el calibre molecular puede funcionar como un mecanismo ancestral de selección de especificidad, aunque deberían existir otros selectores de especificidad para explicar la regulación de genes controlados por ARF. Nuestras estructuras también proponen una explicación interesante para la alta afinidad del ADN, donde el movimiento de la cadena lateral de la His136 (numeración AtARF1) permite una interacción de alta afinidad con ciertas secuencias de ADN. Esta histidina no está presente en los ARF de clase C, lo que representa una diferencia en la afinidad del ADN entre las clases A/B y C.

Nuestro análisis estructural también se amplía al dominio auxiliar de ARF (**Ancillary Domain**, AD), un subdominio del DBD sin función previa asignada. El pliegue del dominio auxiliar está relacionado con los dominios Tudor, un miembro de la "**Royal Family**" (RF) de lectores de histona metiladas en lisina y arginina. Los dominios auxiliares compartían elementos con múltiples subfamilias de RF, lo que nos llevó a comparar la AD con proteínas de RF conocidas. En el **capítulo 4** revisamos la clasificación de la Royal Family, resumiendo los elementos necesarios para que una proteína sea parte de esta superfamilia. El trabajo presentado en el capítulo 4 permitió analizar si los ARF cumplen con las características de la familia de RF. Demostramos en el **capítulo 5** que el ARF-AD retiene todos los elementos estructurales para ser un lector funcional de modificaciones postraduccionales de histonas, y que este dominio no está relacionado en secuencia con ninguna

otra proteína que no sea ARF o proteínas vegetales. Hemos visto que el sitio de unión en los ARF siempre está cubierto por un aminoácido básico, lo que evitaría la unión de moléculas al ARF-AD. Además, encontramos que el pliegue del dominio de dimerización (DD) se asemeja al pliegue de muchos miembros de la superfamilia RF, con un sitio de unión no funcional. En otros dominios de RF, como en el Tudor y cromodominio extendidos, los dos módulos similares a RF proporcionan diferentes funcionalidades. En los ARF, el DD agregaría capacidad de dimerización al AD. No fue posible una clasificación clara de los dominios auxiliares en las familias de proteínas de RF existentes, ya que los dominios auxiliares de ARF compartían elementos con múltiples subfamilias de RF. En consecuencia, proponemos que los ARF constituyen una nueva familia dentro de la RF y acuñamos el término "dominio Steward" para referirnos a los dominios combinados similares a RF que se encuentran en el DD y AD de los ARF. En el **capítulo 6**, demostramos que AtARF1-DBD interactúa con varios péptidos de histonas y que el pH afecta la interacción. Además, los análisis de arrays nos mostraron un abanico de interacciones de AtARF1-DBD con péptidos derivados de histonas modificadas, lo que sugiere una funcionalidad de módulo de lectura de histonas *in vivo*, que seleccionaría genes en presencia de auxinas en función del estado epigenético del nucleosoma. Esto puede representar la pieza que falta en el rompecabezas de la regulación ARF.

Finalmente, en el **capítulo 7** discutimos y resumimos toda la información recopilada en este trabajo y se presentan las direcciones y consideraciones futuras.



| | | |
|------------|---|-----------|
| 1 | Introduction | 19 |
| 1.1 | Scope of this thesis | 22 |
| 1 | Molecular callipers | |
| 2 | Structural studies on the ARF-DNA affinity | 27 |
| 2.1 | Abstract | 27 |
| 2.2 | Introduction | 28 |
| 2.3 | Results | 31 |
| 2.3.1 | DBD architecture is conserved since <i>Marchantiophyta</i> and <i>Brassicaceae</i> divergence | 31 |
| 2.3.2 | Structural insights into the calliper model | 33 |
| 2.3.3 | Structural determinants of B3 specificity | 36 |
| 2.4 | Discussion | 38 |

| | | |
|------------|---|-----------|
| 3 | Testing the calliper model in solution | 43 |
| 3.1 | Abstract | 43 |
| 3.2 | Introduction | 43 |
| 3.3 | Results | 45 |
| 3.3.1 | ARFs show sequence and spacing dependent affinity <i>in vitro</i> | 45 |
| 3.3.2 | DNA binding drives ARF dimerization | 47 |
| 3.3.3 | The ARF-DBD alone is enough to promote ARF-ARF heterodimerization | 50 |
| 3.4 | Discussion | 52 |

II

Ancillary Domain

| | | |
|------------|--|-----------|
| 4 | The Royal Family of methylation readers | 57 |
| 4.1 | Abstract | 57 |
| 4.2 | Introduction | 57 |
| 4.3 | The Royal Family | 59 |
| 4.3.1 | Chromo-like domains | 60 |
| 4.3.2 | Tudor-like domains | 64 |
| 4.3.3 | Non-RF Histone methyllysine readers | 71 |
| 4.4 | Discussion | 72 |
| 5 | In search of an Ancillary Domain function | 77 |
| 5.1 | Abstract | 77 |
| 5.2 | Introduction | 77 |
| 5.3 | Results | 78 |
| 5.3.1 | ARF Ancillary Domain, founding member of a plant-specific Royal Family-like domain | 78 |
| 5.3.2 | ARF Tudor-like Dimerization Domain | 79 |

| | | |
|------------|--|------------|
| 5.3.3 | ARF-DBD Tudor-like Ancillary Domain | 80 |
| 5.3.4 | The ARF-AD Hydrophobic Cage is structurally related to the Royal Family HC . | 81 |
| 5.3.5 | Putative ARF cage residues are conserved during evolution | 82 |
| 5.3.6 | Sequence homology of the ARF ancillary domain | 84 |
| 5.3.7 | The Hydrophobic Cage entrance is regulated by a basic residue | 85 |
| 5.3.8 | The Ancillary Domain shares structure with most Royal Family domains | 90 |
| 5.4 | Discussion | 92 |
| 6 | Steward Domain: The ARF epigenetic link | 95 |
| 6.1 | Abstract | 95 |
| 6.2 | Introduction | 95 |
| 6.3 | Results | 97 |
| 6.3.1 | The Ancillary Domain alone is not sufficient for PTM binding | 97 |
| 6.3.2 | AtARF1-DBD as an effective histone PTM reader | 98 |
| 6.3.3 | AtARF1-DBD binds RF substrates with low micromolar affinity | 102 |
| 6.4 | Discussion | 104 |
| 6.4.1 | Steward domains as substitutes for histone PTMs antibodies | 106 |

III

Final considerations

| | | |
|------------|--|------------|
| 7 | Discussion, conclusions and future perspectives | 111 |
| 7.1 | General Discussion | 111 |
| 7.1.1 | The B3 conformational freedom is an ancestral mechanism of gene selection | 112 |
| 7.1.2 | Solution studies agree with crystallography findings | 112 |
| 7.1.3 | Similarity between ARF-DBD may be explained by their <i>in vivo</i> activity | 113 |
| 7.1.4 | ARF Steward domain as selector of ARF gene specificity | 114 |
| 7.1.5 | Model of signal integration by ARFs | 115 |
| 7.2 | Conclusions | 117 |

| | | |
|-----------|---|------------|
| 7.3 | Future perspectives | 117 |
| 8 | Materials and Methods | 119 |
| 8.1 | Protein expression and purification | 119 |
| 8.2 | Preparation of dsDNA | 120 |
| 8.3 | Crystallography | 120 |
| 8.3.1 | MpARF2-DBD:21ds C_2 | 120 |
| 8.3.2 | MpARF2-DBD:21ds $I_212_12_1$ | 121 |
| 8.3.3 | MpARF2-DBD:ER7 $I_212_12_1$ | 121 |
| 8.3.4 | AtARF1-DBD:21ds P_21 | 121 |
| 8.3.5 | Structure solution | 122 |
| 8.4 | RMSD values and distance calculation | 122 |
| 8.5 | Analytical SEC | 123 |
| 8.6 | SAXS analysis | 123 |
| 8.7 | Dot-Blot assays | 123 |
| 8.8 | Structural superposition analysis | 124 |
| 8.9 | Peptide microarray assays | 124 |
| 8.10 | Fluorescence anisotropy assays | 125 |
| 8.11 | <i>Marchantia polymorpha</i> plant extracts | 125 |
| 8.12 | Pull-down assays | 126 |
| 8.13 | Overlay blots | 126 |
| 9 | Supplementary information | 127 |
| 10 | Bibliography | 137 |
| | Publications | 155 |

| | |
|--|------------|
| Deposited structures (PDBs) | 156 |
|--|------------|



1. Introduction

All living organisms share a common mechanism of information storage and transmission to produce functional units for life development and maintenance. The complete instruction set specifying the life program of an organism is stored in its genome, which must be translated to the proteins that will perform the cellular functions [1, 2]. Transcription is the process by which a gene, a functional piece of this genomic code, is read and this is a highly regulated process that results in the production of several copies of messenger RNA to be translated into proteins [3]. The regulation of transcription will determine the set of genes that are expressed at any given moment and for a certain cell type, defining the fate of cellular variety [3, 4, 5]. Transcriptional regulation starts by selecting the set of genes that must be expressed or silenced and the expression levels required for the correct cellular function in a determined time and location. Transcription Factors (TFs), along with other proteins, are the actors responsible for this selection mechanism. TFs are proteins with the ability to bind specific DNA sequences [6] located on the promoters of the genes under the control of a determined TF. TFs have the ability to discriminate among similar DNA sequences [7]. In addition, TFs also function as an anchor point for other cofactors, behaving as recruiters to the DNA binding site of proteins that will allow the expression or silencing of genes [4, 7]. TFs can be regulated at different levels, and the factors that determine the output produced at each of the regulation steps contribute to the complexity of gene regulation. For example, the presence of a certain TF can activate or block the expression of a gene. Regulating the synthesis or cellular localization of a TF will affect the expression of genes regulated by the TF [8]. Certain TFs need activation, or other cofactors to perform their function, and thus variations in the amount of activator or interacting protein can fine-tune gene expression [7]. Finally, the accessibility to the DNA site by the TF is fundamental for recognition. A clear example is the effect of histones on DNA compaction and decompaction, where epigenetic marks play a key role in the definition of active and silenced DNA regions [9]. An accessible DNA site will promote the binding of TF, facilitating the expression of the genes located on this region [10].

Overall, TFs integrate all the input received from both the extracellular environment and intracellular signals. A better understanding of TFs is essential for understanding gene expression

regulation. An example of tight regulation of gene expression is observed during embryogenesis, where it serves to ensure that a specific cell type is produced in a specific place. For this regulation, hormones like Retinoic Acid in animals and phytohormones like auxin in plants are essential for the establishment of a symmetry axis, crucial for the correct development of the different structures of the embryo [11, 12, 13]. Regulation failures will lead to fatal developmental errors. This regulation is not exclusive to development, as the organism will require proper gene regulation during all its life to maintain cell identity and control biological processes. Defective gene regulation can in addition lead to several abnormalities, such as cancers and autoimmunity in humans, or fruit discoloration and altered sugar content, growth and developmental abnormalities in plants [4, 14, 15, 16, 17].

The interest towards plant genetic regulation has increased in the present context of global warming and growing population, where strategies to increase the yield for the limited availability of crop area are required to overcome food shortage [18]. Arable weeds are the major biotic cause of crop losses as they compete with cultures for light, nutrients and water, resulting in yield losses of almost one third of the worldwide production, with an economic cost greater than 26 billion US\$ in the USA alone [19, 20]. Since the first herbicides were marketed in the 1940s, weed control has been more effective compared to the previous laborious methods, which were based on the manual extraction of weeds from the soil. However, the intense selective pressure exerted by continuous herbicide application led to the adaptive evolution of weed species, as the less sensitive individuals have a reproductive advantage in weed populations subject to herbicide treatment. The reproductive advantage translates to further spread of the genes that promote the resistance through generations until the whole weed population shifts to a high frequency of herbicide-resistant individuals. According to the Herbicide Resistance Action Committee (HRAC) classification [19], there are 18 classes of herbicides [19, 20]. Each class represents different target sites, which is either a mode of action or the way in which the herbicide controls the susceptible plants. All the herbicides whose mode of action is unknown are grouped in class Z. Classes A, K3, and N target fatty acid biosynthesis; B, G and H affect different pathways of amino acid synthesis; C, D, E, F act at different steps of the photosynthesis; I class targets tetrahydrofolate biosynthesis; K1 and K2 inhibit microtubule polymerization; Class L herbicides inhibit cell wall biosynthesis; and class M target ATP biosynthesis. Finally, classes O and P are targeting auxin signalling, acting as artificial auxins stimulating Transport Inhibitor Response protein 1 (TIR1), and impairing auxin transport, respectively [19, 20].

Auxin is the main growth phytohormone, discovered during the 1930s, which regulates all critical aspects of plant development from early embryogenesis to fruit ripening, organogenesis, cell division, cell expansion and differentiation, root initiation and others [21, 22, 23]. The principal natural auxin in higher plants is indole-3-acetic acid (IAA). Cells that sense auxin can response fast to it through non –transcriptional cellular responses, but many of the developmental responses triggered by auxin are mediated by gene expression regulation [24]. Auxin controls gene expression through a rather simple pathway, the Nuclear Auxin Pathway (NAP). NAP consists of three components, the auxin co-receptors F-box TRANSPORT INHIBITOR RESPONSE 1/AUXIN SIGNALING F-BOX PROTEIN (TIR1/AFB), the Auxin/INDOLE-3-ACETIC ACID (Aux/IAA) transcriptional repressors and the AUXIN RESPONSE FACTOR (ARF) transcription factors [25]. Although NAP seems too simple to explain all the variable genetic responses to auxin, the three members belong to multigene families. Thus, in *Arabidopsis thaliana* there exist 6, 29 and 23 genes coding for TIR1/AFB, Aux/IAA and ARF, respectively [26, 27]. In addition, differential expression patterns of each of these genes at different tissues increase the complexity of responses to auxin presence [26, 28, 29].

In absence or at low intracellular auxin concentrations, Aux/IAA is bound to and represses ARF transcription factors [30]. The first ARF was discovered in 1997 as a transcription factor from *Arabidopsis thaliana* (AtARF1) that bound to the sequence TGTCTC in AuxREs, containing an amino-terminal DNA-binding domain [31]. Since then, a large amount of knowledge on the AtARF gene redundancy, the structure and the physiological function has accumulated. ARFs are formed by at least two different domains: an N-terminal domain comprising the DNA-binding domain (DBD), which is responsible for the DNA-binding ability of ARFs; and the middle region (MR), unstructured and with the ability to recruit partners to the ARF site [32, 33]. Most ARFs also contain a C-Terminal Phox and Bem 1 (PB1) protein-protein interaction domain, essential for the interaction between ARFs and Aux/IAA proteins and probably for ARF-ARF oligomerization [24, 30, 34, 35, 36]. The PB1 domain is where Aux/IAA docks and which prevents ARF function [35]. As auxin concentrations increase, auxin binds to TIR1/AFB auxin receptor, creating a binding site for the Aux/IAA transcriptional coregulators. The complex formation of TIR1/AFB, auxin and Aux/IAA will trigger ubiquitin-mediated degradation of the Aux/IAA proteins via the proteasome, releasing ARF action [20, 22, 24, 25, 37] (Figure 1.1).

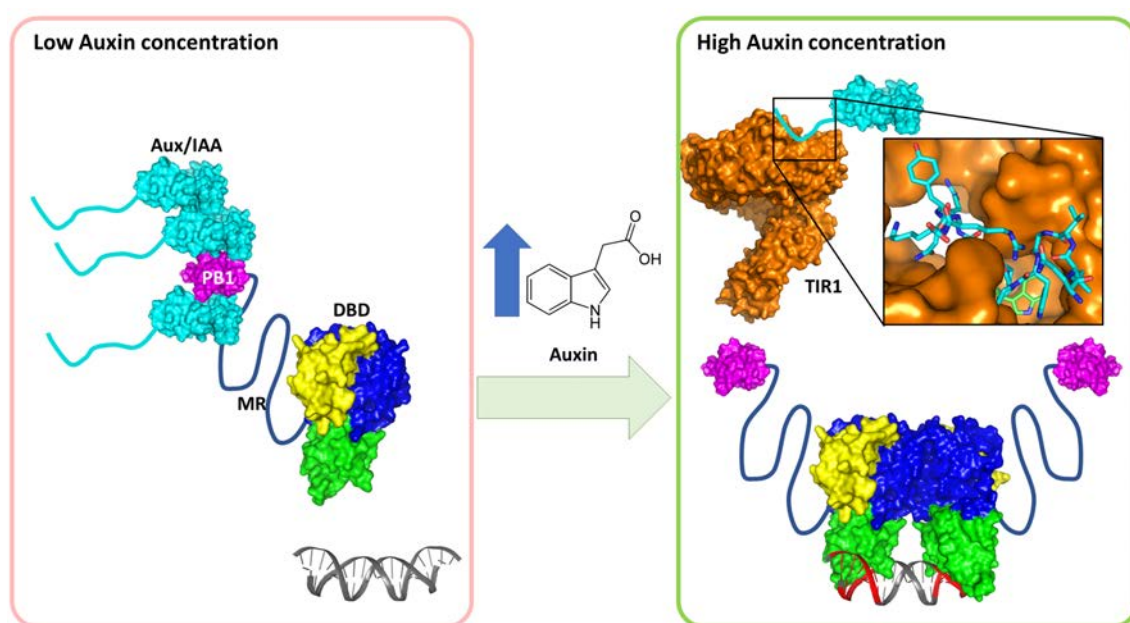


Figure 1.1: Schematic representation of ARF activation on auxin concentration increase. Under low auxin concentration, Aux/IAA proteins are locking ARFs by interacting with their PB1 domains with the PB1 domain present in ARF C-terminus. In high auxin concentration, auxin is sensed by TIR1, which is able to recognize unstructured tails in Aux/IAA thanks to the mediation of auxin (Insert, top view of Aux/IAA:TIR1:auxin complex, auxin is shown as green sticks, Aux/IAA unstructured tail as blue sticks, TIR1 surface in orange). This interaction subsequently produces the polyubiquitination of Aux/IAA and the proteosomal degradation, leaving ARFs free to perform their action.

ARFs have been extensively studied as they are believed to be the most important actors in transcriptional response to auxin due to their N-terminal DBD specificity for Auxin-Response DNA Elements (AuxRE) and due to their activation in the last steps of the auxin signalling pathway [33]. AuxREs contain the canonical core sequence TGTCNN [38], *e.g.*, the TGTCGG motif [39], which has a high affinity for AtARF1. AuxREs are located in the promoters of auxin-regulated genes and ARF binding will either activate or repress transcription of the target genes [33, 40]. The C-terminal

region of ARFs also attracted the attention of researchers due to its ability to interact with other proteins, forming a regulatory spot for auxin response. This interest led to several structural studies [30, 34, 37] that demonstrated the importance of this domain. On the other hand, the middle region is much less studied, but it is known that is necessary for their function despite being almost fully unstructured [41]. Crystallographic structures of the DNA Binding Domain showed that the ARF-DBD is in turn also formed by three distinct structural domains: a B3 DNA-binding subdomain embedded in the middle of a Dimerization Domain (DD) that allows DBDs to dimerize, and the last 80 amino acids fold into a small subdomain, the Ancillary Domain, with no attributed function [39]. The DBD is also important for the nuclear localization of the transcription factor, as nuclear localization sequences are found within the DBD [31, 42].

Land plant ARFs are phylogenetically grouped into three conserved classes, named A, B and C [28, 43]. Transactivation assays classifies class A ARFs as activators, while class B members are considered repressors [44, 45]. It has been suggested that the composition of the MR may be linked with the ability to activate or repress transcription [46, 47, 48], due to selection of the recruiting partners depending on the MR sequence [48]. A MR rich in leucine, serine, and glutamine residues is associated with an activating ARF, while a MR enriched in proline, glycine, threonine and serine residues is present in repressing ARFs [49]. Based on the amino acid composition of the MR, Class C ARFs are generally believed to be transcriptional repressors, but class C ARFs are the most diverging group of the ARF family [50]. Thus, the MR determines whether transcription is activated or repressed, but the mechanism of DNA selection is less clear. The crystal structures of AtARF1 and AtARF5 DBDs highlighted that the structure and the amino acid residues responsible of the interaction with the DNA were almost conserved between ARFs [39]. This left the question open on how ARFs with a highly conserved DBD are able to select different genes for expression or repression. A better understanding of how ARFs discriminate their target genes is required, as the current knowledge is unable to explain the capability of the rather simple auxin molecule to control and regulate all the different auxin-responsive genes. Improved understanding would help in the comprehension of the pathway, the ability to tune the auxin pathway and ultimately, regulate plant growth and development.

1.1 Scope of this thesis

Auxin Response Factors comprise a family of transcription factors in charge of regulating gene expression in response to auxin. Despite the knowledge accumulated on the Nuclear Auxin Pathway and ARFs during the recent years, there is still little information on how this rather simple pathway can control all the physiological processes regulated by auxins.

In this work, we focus our efforts on understanding the structural and molecular basis of gene selection by ARFs. For this purpose, we study ARFs from the model organisms *Arabidopsis thaliana* and *Marchantia polymorpha*. *M. polymorpha* is considered a distant relative of *A. thaliana* that may resemble the ancestral state of land plants, where findings indicate that it established the basic auxin nuclear pathway, containing one protein for each function: TIR1/AFB, AUX/IAA and one member for each of the three classes of ARF transcription factors [26]. Our work aims at finding structural features to understand the specificity for the gene regulation.

Part One of this thesis comprises **Chapters 2 and 3** and is devoted to test the “molecular calliper” hypothesis originally proposed by Boer and co-workers [39], which tries to explain how different ARFs can specifically recognize their target genes regardless of sharing similar B3 domains, based the recognition on AuxRE spacing:

Chapter 2 reports the crystal structures of the DNA Binding Domain of ARF2 from *Marchantia polymorpha* in complex with the canonical ER7 and high affinity 21ds sequences. The structure of the DNA Binding Domain of ARF1 from *Arabidopsis thaliana* in complex with the high affinity 21ds sequence is also reported. This chapter focusses on the comparison between *Arabidopsis* and *Marchantia* ARFs and between class A and B ARFs. A mechanism of AuxRE recognition with high affinity involving His136 (AtARF1 numbering) alternate conformations is proposed. This histidine, not present in class C ARFs, is the only, yet remarkable, difference found among ARF DNA binding residues.

Chapter 3 thoroughly tests the previously proposed calliper model in solution by means of Analytical Size Exclusion Chromatography and Small-Angle X-Ray Scattering. For these analyses we examined the behaviour of 4 different members of the ARF family from *Arabidopsis thaliana* and *Marchantia polymorpha*, representing the 3 classes, and their interactions with several AuxREs and spacing. This allowed us to compare evolutionary distant homologues. Finally, we also included Dot-Blot analyses to demonstrate the ARF-DBD ability to homo/heterodimerize in absence of the C-terminal PB1 domain. This feature is conserved between *Arabidopsis* and *Marchantia* proteins.

Part Two of this work comprises **Chapters 4, 5 and 6** and is dedicated to test the hypothesis of ARF gene regulation by specific ARF interaction with histone posttranslational modifications. This hypothesis is based on the analysis of the structures presented in **Part One** of this work. These structures suggest a protein-protein interaction module located in the hitherto not so well studied Ancillary Domain of the ARF-DBD. Interactomic studies on ARF Ancillary Domains are conducted in order to understand the selection of genes based on epigenetics:

Chapter 4 studies the original connection between the ARF-AD structure and the structure of the Tudor domain in more detail. As Tudor domains are Royal Family members, this chapter reviews and discusses the current knowledge on the Royal Family members with the aim of thoroughly establishing the relationship between the ARF-AD and the Royal family.

Chapter 5 analyses the present knowledge on the ARF-DBD Ancillary Domain, using in-depth bioinformatic and structural analyses of all the available ARF structures. Evidence for a functional Royal Family-like Ancillary Domain is presented and the implications of these findings on ARF gene regulation are discussed. Finally, ARF-ADs are proposed to constitute a novel plant-specific family within the Royal Family of domains.

Chapter 6 reports experimental findings of *in vitro* studies to determine the function of the ARF Ancillary Domain. The applied techniques include a histone hybridization array to analyse affinity of ARF-DBD for histone posttranslational modifications and fluorescence anisotropy assays to determine binding affinities of ARF-DBD for the peptides.

Part three discusses in **Chapter 7** the main findings of this work, and summarizes the conclusions of the results. Future directions to continue on ARF signalling research are also suggested.



Molecular callipers

2 Structural studies on the ARF-DNA affinity **27**

- 2.1 Abstract
- 2.2 Introduction
- 2.3 Results
- 2.4 Discussion

3 Testing the calliper model in solution . . . **43**

- 3.1 Abstract
- 3.2 Introduction
- 3.3 Results
- 3.4 Discussion



2. Structural studies on the ARF-DNA affinity

*Part of this work has been published as **Design principles of a minimal auxin response system** [28] and as **Architecture of DNA elements mediating ARF transcription factor binding and auxin-responsive gene expression in Arabidopsis** [51]*

2.1 Abstract

Many important processes in plants as growth and organ development are controlled by the phytohormone auxin. Plant cells regulate transcription of genes under auxin control via the Nuclear Auxin Pathway (NAP), which ultimately relies on Auxin Response Factors (ARFs) for specific gene selection. Although a significant amount of knowledge on NAP has been accumulated in the past decades, little is known on how ARFs specifically select target genes. The first structures of *Arabidopsis thaliana* ARF DNA Binding Domains (ARF-DBD) shed light on the DNA recognition mechanism, but the high conservation of the amino acids responsible of DNA binding raised questions on the specific promoter selection mechanism by ARFs. This led to the proposal of the molecular calliper model, where the spacing between palindromic DNA sequences determines which ARF is allowed dimerize and bind DNA effectively. In this chapter, we report a structural comparison between A and B classes from *Marchantia polymorpha* and *Arabidopsis thaliana* ARFs to deepen our understanding of the molecular calliper model. We report new *Marchantia polymorpha* structures that allow comparing the evolution of ARF-DBD since *Marchantiophyta* and *Brassicaceae* divergence. Although our data agrees with the contribution of molecular calliper model in DNA recognition, it can only explain part of the complexity of NAP in higher plants. This implies that additional regulatory circuits are necessary to support the specificity observed in organisms with multiple members on each class. We believe that this model of DNA recognition is an ancestral mechanism of gene selection and regulation.

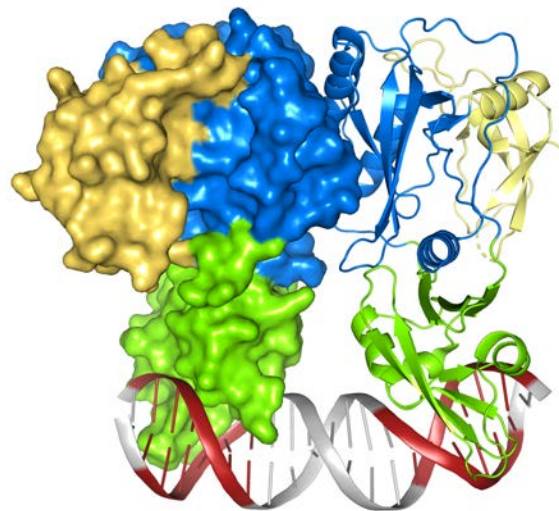
2.2 Introduction

The Auxin Response Factor family of transcription factors regulates gene expression in response to auxin by altering chromatin topology [46, 48]. Gene repression involves recruitment of histone deacetylases by TOPLESS corepressor (TPL) and Aux/IAA repressor, whereas recruitment of SPLOYED/BRAHMA (SYD/BRM) will mark chromatin for active transcription [24]. The ability of an ARF to recruit one or another type of coregulator will determine the state of the genes under ARF control. However, until now few interactions between coregulators and ARFs has been identified [24].

Auxin Response Factors generally consist of several conserved domains, where each domain conveys a determined function. These domains form blocks that are found in other proteins, and *in vitro* assays demonstrate that their folding is independent from the folding of other protein regions [39, 52], which suggests that ARF domains are functionally autonomous. ARF sequences contain an N-terminal **DNA binding domain (DBD)**, followed by a **middle region (MR)**. The MR is mostly unstructured and is variable in length. The MR is enriched in either glutamine or serine, which is correlated with the ability of an ARF to activate or repress expression of target genes, respectively. Finally, C-terminal to MR a **Phox/Bem1p (PB1)** protein-protein interaction domain (Domain III/IV) is found in most ARFs [37, 53]. The crystal and solution structures of PB1 domains found in ARFs and Aux/IAA proteins consist of two subdomains, corresponding to subdomain III and subdomain IV. The N-Terminal subdomain III consists of an antiparallel β -sheet of strands $\beta 1$ and $\beta 2$ and helix $\alpha 1$. The C-Terminal subdomain IV is also formed by an antiparallel β -sheet of 3 β -strands $\beta 3$ to $\beta 5$ and two α -helices, $\alpha 2$ and $\alpha 3$. The PB1 domain can oligomerize thanks to the distribution of positive and negative charges in the two faces of the protein. The C-terminal PB1 domain of ARFs mediates the interaction between the ARF transcription factor and other proteins [39]. Among the known interactors, the Aux/IAA inhibitors repress ARF function under low auxin concentration, being degraded by proteasome action in presence of auxin, thereby releasing the ARFs. It has also been proposed that ARFs can oligomerize through this domain [54]. Most class B and C ARF members have limited interaction capacities with Aux/IAAs. Due to this limited interaction, class B and C ARFs were proposed to regulate auxin transcriptional responses in an auxin-independent manner, possibly by competition with class A for binding sites [55].

The first crystal structures of ARF-DBD revealed the modularity of the DNA Binding Domain as it consists of three subdomains [39]. The first subdomain is known as **Dimerization Domain**, which was found to allow the homodimerization of ARF-DBD with no other requirements. The ARF-DBD dimerization is critical for its function, as mutants at amino acids located in the dimerization interface disrupting ARF dimerization resulted in AtARF5 dysfunction *in vivo* [39]. Embedded in this domain a **plant specific B3 subdomain** is found, which is responsible for the DNA binding properties of ARF-DBD. The 80 C-terminal residues of ARF-DBD form an extra domain known as **Ancillary Domain (AD)** [26, 32, 33, 35, 39, 43, 56] or Flanking Domain [33, 35, 38], with no attributed function yet [32, 33, 35, 43] (Figure 2.1).

Interestingly, the B3 superfamily of DNA Binding Domains is only found in plants, and are surprisingly similar to the noncatalytic DBDs of certain restriction endonucleases from *Escherichia coli* and *Bacillus firmus* [57, 58]. All B3 domains consist of 110 amino acids that fold in a seven-stranded open β barrel structure and two α -helices placed at the ends of the barrel. Aromatic residues in the α -helices interact with the hydrophobic core of the barrel, stabilizing the structure [39, 58, 59]. The B3 domain present in ARFs is also found in **ABI3** (Abscisic Acid insensitive 3, B3ABI3) and **RAV** (Related to ABI/VP1, B3RAV) plant transcription factors but with different DNA binding specificities: the recognition sequence motifs identified for the ARF family members,



ER7: TGTCTCCCTTTGGGAGACA
21ds: TGTCGGCGATTCGCCGACA

Figure 2.1: DBD Subdomain organization in solved ARF structures. The DBD dimerizes through Dimerization Domain (blue) to interact with AuxRE (DNA sequences coloured in red) as a dimer by means of the B3 domains (green). The Ancillary Domain is shown in yellow. One of the DBDs forming the dimer is shown as surface, while the other is shown as cartoon. The ER7 and 21ds sequences used for the determination of the structures of the ARF-DNA complexes are shown, where the AuxRE sequence is coloured in red and the seven nucleotide spacing in black.

the canonical 5'-TGTCTC-3' AuxRE sequence (**ER7**, 5'-TGTCTC**c**ctttg**G**AGACA-3') or the non-palindromic **TMO5** binding sequence (5'-GGTCTC**t**ggtc**g**GCAGA-3'), differ from those of the ABI3/VP1 members of the LAV family (5'-CATGCA-3'), and of the B3 domain of RAV family members (5'-CACCTG-3') [57, 58, 60]. Analysis of the binding affinity of ARF-DBDs by protein binding microarrays showed higher binding affinities for the 5'-TGTCGG-3' (**21ds**, 5'-TGTCGG**c**gat**t**c**C**CGACA-3') sequence, but the mechanism for higher affinity is not understood.

Contrary to vertebrates and other species, cytosine methylation at carbon five can occur in plants on any cytosine, regardless of the sequence [61]. This results in a high rate of methylation in plants. The canonical ER7 and high affinity 21ds AuxRE sequences contain possible methylation sites, which may impact ARF signalling. Methylation of DNA is one important epigenetic change which influences gene expression regulation [62]. According to reports on genomic DNA methylation, the methylation rate in *Arabidopsis* is extremely high, as approximately 14% of cytosines are methylated [63] while in other organisms, like *Homo sapiens* and *Mus musculus*, the methylation rate is 5.8 and 7.6%, respectively [64]. Certain organisms, like insects and yeast, present a methylation rate which is virtually zero [63]. The *Arabidopsis* genome contains methylation at 24% of CG sites, 6.7% of CHG and 1.7% of CHH, being H a non G base [61]. According to these data, the ER7, TMO5 and 21ds sequences possess a putative methylation site in C₄. In the ER sequence, TGTCTC/GAGACA forms a CHH site, with low methylation probability (1.7%). In the case of 21ds, the TGTCGG/CCGACA sequence forms a CG site, a site with high probability of methylation (24%), as in the case of TMO5, where a CHH (GGTCTC/GAGACC) and CG (TCGACA/TGTCGA) sites can be found. This suggests that DNA methylation in AuxREs plays a role in gene expression regulation by ARFs. Studies of B3 DNA recognition of methylated sequences would help to further understand the binding mechanisms.

possible orientations: **Direct Repeat (DR)**, **Everted Repeat (ER)** and **Inverted Repeat (IR)** [38]. DBD dimerization led to the proposal of the “**molecular calliper**” hypothesis, in which ARFs structural restraints on the spacing of B3 domains would allow different ARFs to bind uniquely spaced palindromic motifs [37, 39, 44]. Since the ARF calliper model was proposed, other plant transcription factors as LOB domain family of zinc-finger transcription factors have been proposed to follow a similar mechanism of gene selection [65]. The divergence observed in loops connecting B3 and DD in AtARF1-DBD and AtARF5-DBD structures indicate that DBD intersubdomain flexibility can be responsible of the selection of distinctive AuxRE spacing [39]. *In vivo* results with AtARF5 support this hypothesis [66, 67] and suggests an AtARF1-DBD narrower range of preferences than AtARF5-DBD [38] despite their similar AuxRE binding affinity [56]. Promoter analysis in genes regulated by different ARFs also show distinct spacing patterns of AuxREs [32].

In this chapter we present the first crystallographic structure of *Marchantia polymorpha* ARF2 in complex with the high affinity **21ds** sequence [28] found for AtARF1-DBD [39]. This first low resolution structure revealed that the ARF-DBD architecture is conserved since *Marchantiophyta* and *Brassicaceae* divergence, more than 400 million years ago. We also report two more MpARF2-DBD structures in complex with **ER7** and **21ds** sequences at higher resolution and the structure of *Arabidopsis thaliana* ARF1 in complex with 21ds sequence [51]. These structures were useful for deepening our understanding of the structural restraints of ARF-DBDs to accommodate different AuxRE spacings, with the aim to better understand and refine the molecular callipers hypothesis. With this purpose, we looked at the changes in the relative orientation of B3 and DD domains, measured by the B3/DD twist angle, and we compared the distance between B3s in the Apo and DNA-bound structures, which suggest a preference of AtARF1 for higher spacing than MpARF2 and AtARF5. Finally, we focus on the structural determinants of B3 binding preference in AtARF1 for 21ds compared with ER7. We suggest that the main reason for the higher affinity is a change in conformation of H136, which is substituted for a glycine in class C ARFs, to a position that is establishing more contacts with DNA bases when bound to 21ds.

2.3 Results

2.3.1 DBD architecture is conserved since *Marchantiophyta* and *Brassicaceae* divergence

The first structure of an ARF-DBD from *Marchantia polymorpha*, an *Arabidopsis* distant relative that may resemble the ancestral state, was solved in complex with the high affinity 21ds DNA at a resolution of 2.96 Å (PDB: 6SDG, C2 space group, **LR-Mp2:21ds**) [28]. This structure revealed that the overall structure is very similar to the *Arabidopsis thaliana* structures already reported [39], showing that the DBD architecture has remained conserved since *Marchantiophyta* and *Brassicaceae* divergence, more than 400 million years ago. The DBD in both *Arabidopsis* and *Marchantia* structures is formed by a Dimerization Domain, where a B3 DNA interaction domain is embedded, and an Ancillary Domain, with no clear function [32, 33, 35, 43] (Figure 2.3).

The MpARF2-DBD is bound to 21ds DNA as dimer, where dimerization occurs trough $\alpha 6$ – *helix*, similarly to what is observed in the previous ApoAtARF1 (PDB: 4LDY, **ApoAt1**), AtARF1-DBD:ER7 (PDB: 4LDX, **At1:ER7**) and Apo AtARF5-DBD structures (PDB: 4LDU, **ApoAt5**) (Figure 2.3A). Due to the low resolution of LR-Mp2:21ds structure, some secondary structural elements were not automatically recognized by modelling software and were manually imposed by similarity with At1:ER7 (Figure 2.3B). As observed in the AtARF structures, $\alpha 1$ – *helix* is

of ARFs analysed, 3 α -helix-like are found, $\alpha 4$ being located at the outside part of the dimer and $\alpha 3$ and $\alpha 5$ facing inside, to the DNA spacer. In RAV B3, the connecting loop between $\beta 4$ and $\beta 5$ is shorter and has an α helix-like conformation, but this loop in ARFs is longer and is able to form the third α -helix, which is located in the major groove of the interacting DNA (Figure 2.4). The B3 structure is highly conserved, where the main differences are localized on the three loops of variable length connecting $\alpha 3 - \beta 5$, $\alpha 4 - \beta 6$, $\beta 9 - \beta 10$ [59]. The possible formation of this α -helix in B3 domains should be considered for the annotation of structural elements present in other B3 domains.

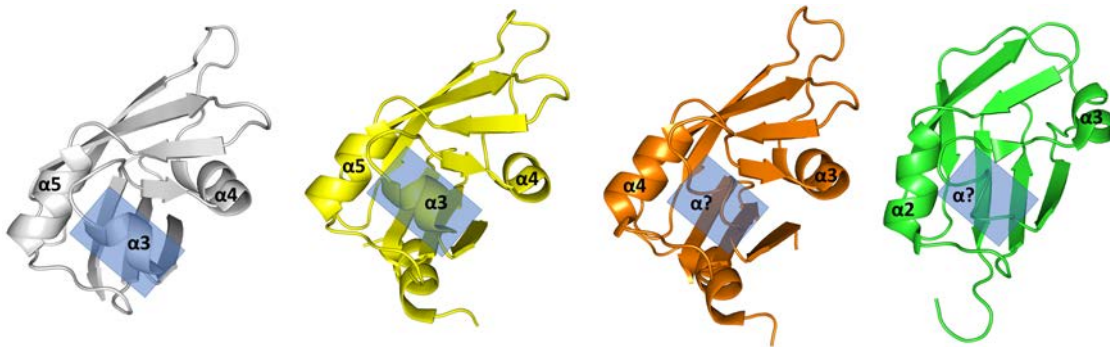


Figure 2.4: B3 domains of Auxin Response Factors contain 3 α helices. B3 domains in ApoAt1 (grey) and LR-Mp2:21ds (yellow) present 3 helices, one of them not previously observed in other B3 domains (blue boxes). This observation can be extended at least to ApoAt5 B3 (orange) and RAV B3 (green).

Regarding the Dimerization Domain, the “Taco” shape observed in ApoAt1 and ApoAt5 structures is also conserved. Although strands $\beta 2$, $\beta 3$ and $\beta 11$ are not completely modelled as β -strands in LR-Mp2:21ds structure, the positions of $C\alpha$ are similar, so probably this is an artefact of the low resolution of the LR-Mp2:21ds structure. Finally, the conformation of the Ancillary Domain remains unaltered between the compared structures. It begins with the $\alpha 7$ helix, followed by 3 antiparallel β strands that form a barrel-like structure covered on its’ side by a 3_10 – helix between strands 16 and 17. The fact that the Ancillary Domain structure is conserved during ARF evolution highlights that it may have a role in ARF function.

2.3.2 Structural insights into the calliper model

Apart from the LR-Mp2:21ds structure described above, we determined three additional structures: MpARF2-DBD in complex with i) **21ds** at higher resolution (2.56 Å, PDB: *Pending assignment*, **HR-Mp2:21ds**) and ii) in complex with **ER7** (2.55 Å, PDB: *Pending assignment*, **Mp2:ER7**) both at space group $I2_12_12_1$, and iii) the structure of AtARF1-DBD in complex with 21ds (1.65 Å, PDB: 6YCQ [51], **At1:21ds**). These structures allow us to further understand the structural and evolutionary aspects of ARFs. The high-resolution structures of MpARF2-DBD presented here show new features of the *Arabidopsis* distant relative MpARF2 that previously were not seen in the LR-Mp2:21ds structure [28]. The higher symmetry found in these new structures contained one MpARF2-DBD monomer and half of the DNA sequence in the asymmetric unit, as the other monomer and DNA half are found by symmetry relationships. The resolution of the At1:21ds structure at 1.65 Å, which is significantly higher than resolution of 2.7 Å of the previously reported At1:ER7 structure, facilitated the construction of some loops (Q228-P233 and E299-K306) that were previously not clearly visible.

The overall structures of At1:21ds and Mp2:21ds are very similar to the respective structures in complex with ER7, with an all residue backbone RMSD of 2.690 Å and 1.027 Å for At1:21ds/ER7 and Mp2:21ds/ER7, respectively (Figure 2.5 A and B), suggesting that the affinity is not determined by DBD conformational changes. Although the overall structure remains practically invariant, a slightly higher curvature can be observed in the DNA for the 21ds structures. This results in a slight contraction of the protein structure along the direction of the main axis of the DNA, which brings the two B3 domains about 0.594 Å closer in average to each other on At1:21ds structure, whereas in HR-Mp2:21ds, the B3 domains are in average 0.653 Å closer to each other.

Overall, the Mp2:21ds structures have relevant displacements of the subdomains when compared with At1:21ds. A structural superposition of the dimerization domain of LR-Mp2:21ds, HR-Mp2:21ds with At1:21ds highlights big differences in the relative position of B3 with respect to the dimerization subdomains (Figure 2.5). In contrast with what could be expected, the B3 position of HR and LR Mp2:21ds complexes are more similar to ApoAt1 (B3 RMSD: 3.354 Å (HR-Mp2:21ds – ApoAt1) (Figure 2.5C), 3.633 Å (LR-Mp2:21ds – ApoAt1)) than to the structure of the At1:21ds complex (B3 RMSD: 6.639 Å (HR-Mp2:21ds – At1:21ds) (Figure 2.5D), 5.302 Å (LR-Mp2:21ds – At1:21ds)). When superposing the DD of the Mp2:DNA structures on ApoAt5 DD instead, the B3 domains are closer, as lower B3 domain RMSDs were obtained: 1.434 Å (ApoAt5 – HR-Mp2:21ds) (Figure 2.5E), 1.432 Å (ApoAt5 – Mp:ER7), and 2.338 Å (ApoAt5 – LR-Mp2:21ds). A similar alignment between ApoAt5 dimerization domain with the At1:21ds dimerization domain resulted in B3 backbone RMSD values of 5.204 Å (Figure 2.5F) and in 6.980 Å in the ApoAt5 – ApoAt1 alignment (Figure 2.5, ??). On the other hand, the change in conformation is remarkable in the structural superposition of ApoAt1 – At1-21ds, with an RMSD value of 7.537 Å (Figure 2.5H).

To quantify the relative B3 displacement along the DNA axis, we measured the distances between C α of equivalent B3 residues from different structures. To perform these pairwise alignments, we selected the first residue of the β 6 and β 8 strands (P160 and R186 in AtARF1) and the last residue of the β 7 (R181 in AtARF1) strand for measuring distances. The choice of these residues was based on the observation that they are conserved between ARFs and that they are located in the area where most of displacement occurs. The measured distances between R223, P202 and R228 in HR-Mp2:21ds with the corresponding R181, P160 and R186 residues in At1:21ds were 6.5 Å, 10.5 Å and 8.9 Å, respectively (8.63 Å on average). In the case of the structural superposition of HR-Mp2:21ds against ApoAt1, the distances between the same residues were reduced to 5.4 Å, 4.6 Å and 4.3 Å (4.77 Å in average). Just as observed before in the B3 all-residue RMSD calculation, the highest similarity in B3 position is found for the HR-Mp2:21ds and ApoAt5 structures. The distances of the residues R223, P202 and R228 in HR-Mp2:21ds with R215, P194 and R220 in AtARF5 are reduced to 0.6 Å, 1.5 Å and 1.1 Å, respectively (1.07 Å on average). The changes in the measured distances may be a consequence of changes in the B3 spacing or to a B3 tilt, pointing outwards with respect to the DNA axis.

During the analysis, we observed that the variation in the measured interatomic distances can contribute to the change in the spacing between the B3s of the dimer along the DNA axis. To quantify the change in B3 spacing, we measured the distance between Gln217 (in AtARF1 numbering), located in the most outer part the B3 domain of each of the monomers forming the homodimer. The distance in ApoAt1 was reduced from 74.9 Å to 56.1 Å upon ER7 binding, while B3 distance in ApoAt5 is 68.9 Å and 54.9 Å in Mp2:ER7. This may indicate a preference of AtARF1 towards longer AuxRE spacing compared to AtARF5, as the B3 domains are spaced further apart in ApoAt1 than in ApoAt5 (Table 2.1).

The changes in B3 position relative to DD may be required for adaptation to different DNA,

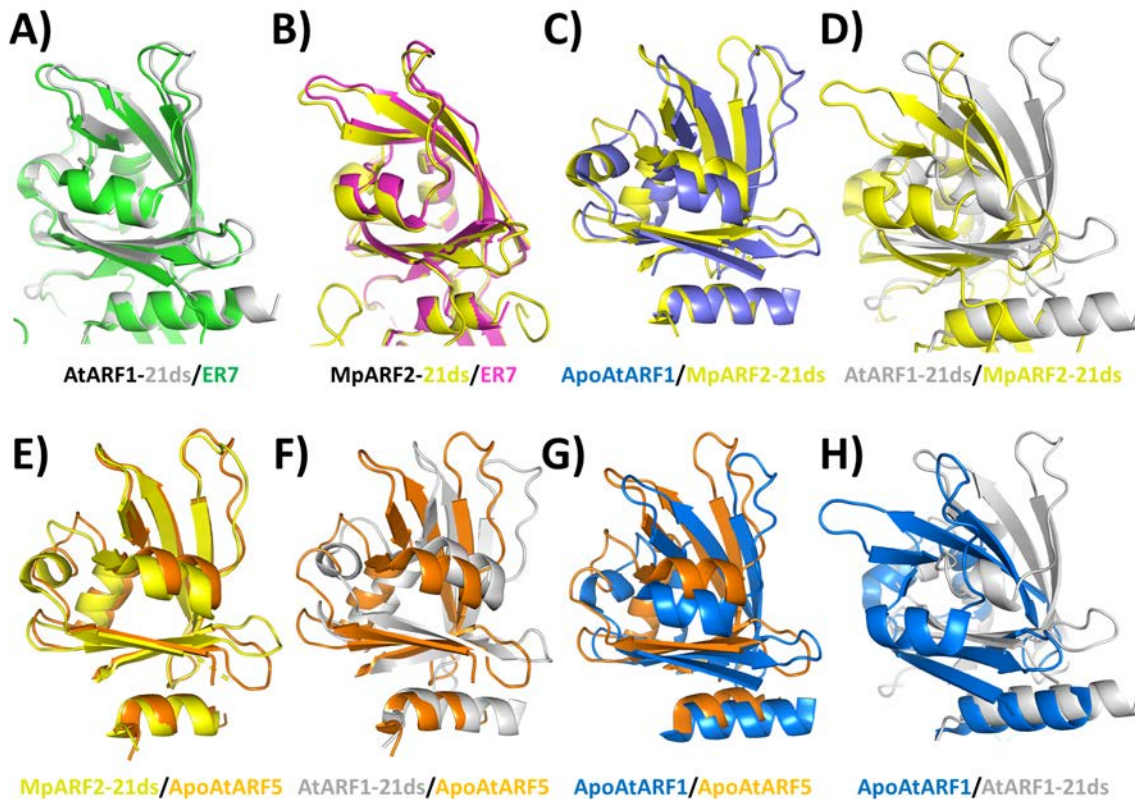


Figure 2.5: Structural comparison between B3 domains of MpARF2, AtARF5 and AtARF1 structures. Cartoon representations of different pairs of B3 domain structures after superposition of the DD for A) At1:21ds (grey) and At1:ER7 (green), B) HR-Mp2:21ds (yellow) and Mp2:ER7 (pink), C) ApoAt1 (blue) and HR-Mp2:21ds (yellow), D) At1:21ds (grey) and HR-Mp2:21ds (yellow), E) HR-Mp2:21ds (yellow) and ApoAt5 (orange), F) ApoAt5 (orange) and At1:21ds structure (grey), G) ApoAt5 (orange) and ApoAt1 (blue), H) ApoAt1 (in blue) and At1:21ds (grey).

as increasing AuxRE spacing also rotates the position of the major groove, which is where the interaction occurs. To quantify the relative movement in the B3 domains relative to the dimerization domains, we structurally superposed the DD of all the structures. Then, we traced a line connecting the C α s of the I1e73 residues (AtARF1 numbering) located in the DD of both monomers of the ARF structure, and, similarly, a line connecting the C α s of the G1n217 residues (AtARF1 numbering) located in both B3 domains. We then projected these two lines on the plane perpendicular to the dimer axis and calculated the angle between the two projections. This angle represents the twist of the B3 domain pair with respect to the DD pair in the plane perpendicular to the dimerization axis and is also a measure of the planarity of the structure. The B3/DD twist angle in the ApoAt5 structure is 23°, whereas in ApoAt1 this angle is 11°. This result shows that ApoAt1 is more planar than ApoAt5. The same measurement performed in At1:ER7 structure resulted in 3° and in At1:21ds structure resulted in 2°, which results in a planarity that is 20° higher with respect to ApoAt5, and 8° between Apo- and DNA-bound AtARF1-DBD structures. In contrast, the analysis of the structures of HR-Mp2:21ds and Mp2:ER7 both resulted in 15°, showing a difference of MpARF2-DNA bound and ApoAtARF5 of 8°.

In general, DNA binding results in more planar structures (B3/DD twist angles close to 0°) and the Apo conformation corresponds to less planar structures (B3/DD twist angles close to 20°). In addition, The B3/DD twist angle in ApoAt1 is smaller compared to that in ApoAt5

Table 2.1: Distance and angle measurements of superposed ARF structures

| B3 RMSDs (Å) | | B3-DD Twist angles (°) | | B3 Distance (Å) | |
|------------------------|-------|------------------------|----|-----------------|------|
| HR-Mp2:21ds – ApoAt1 | 3.354 | ApoAt5 | 23 | ApoAt1 | 74.9 |
| HR-Mp2:21ds – At1:21ds | 6.639 | ApoAt1 | 11 | At1:ER7 | 56.1 |
| ApoAt5 – HR-Mp2:21ds | 1.432 | At1-ER7 | 3 | Mp2:ER7 | 54.9 |
| ApoAt5 – At1:21ds | 5.204 | At1-21ds | 2 | ApoAt5 | 68.9 |
| ApoAt5 – ApoAt1 | 6.98 | HR-Mp2-21ds | 15 | | |
| ApoAt1 – At1-21ds | 7.537 | Mp2-ER7 | 15 | | |

and MpArf2:DNA. Unfortunately, we have not been able to determine AtARF5-DNA bound or ApoMpARF2 structures, which would allow quantifying the magnitude of displacement upon DNA binding for these proteins. Taking into consideration the longer B3 spacing and the higher B3 twist angles observed in AtARF1 structures, the configuration of the B3 domains of the ApoAtARF1 structure fits better, both in orientation and distance, with a DNA sequence in which the AuxREs are spaced further apart.

The analysis of the B3 positions with respect to the DD described above shows that these are similar for MpARF2:DNA and ApoAt5, but different from the DNA-bound structures of AtARF1. This is very interesting and unexpected, as AtARF1 and MpARF2 are classified as class B and AtARF5 as class A based on similarity of the full length sequence. This result shows that in ARFs that are classified as belonging to different classes based on overall sequence homology, the DBDs may have been exchanged between these different classes, contributing to increase the overall complexity of the NAP. To understand the reasons why MpARF2 and AtARF5 B3s are so well structurally superposed we probed the structural determinants of the B3/DD conformations. The structural component with a probable relevance in the position of B3/DD is found in the first α -helix, which is longer in AtARF1. This α -helix is the contact point between DD and B3 domains and acts as a hinge between those domains [39]. The high resolution DNA-bound MpARF2 structures show that the first α -helix of MpARF2 (IDAE LWYACA) and AtARF5 (NSELWHACAG) are both 10 residues long, in contrast to AtARF1, where in At1:21ds it is up to 19 residues long (PGGVLSDALCRELWHACAG) where high sequence conservation is only found between the residues forming the C-terminal part of the AtARF1 that form an helix in MpARF2 and AtARF5. The N-terminal part of the longer AtARF1 α 1-helix is establishing more contacts with β -strands 2 and 3 of the dimerization domain, and with β -strands 4, 9 and 10 of the B3 subdomain of AtARF1, providing more anchor points to the B3 domain in AtARF1, acting as a lock (Figure 2.6). In fact, the formation of the α 1 helical conformation in this region may be a consequence of the longer β -strands 4, 9 and 10 of the B3 subdomain in AtARF1, which in the case of MpARF2 and AtARF5 B3s the extra length in AtARF1 is fold as a loop turn (Figure 2.6). In MpARF2 and AtARF5, the shorter α 1-helix establishes less contacts, leaving a more flexible DD:B3 interface, which results in a structure that is not as planar as in AtARF1. This is in line with the original formulation of the calliper model hypothesis but adds the importance of α 1 – *helix* as a determinant of flexibility, not previously considered.

2.3.3 Structural determinants of B3 specificity

In addition to the contribution of the new MpARF2 and AtARF1 structures to improved understanding of the flexibility of the B3 domain, the structure solution of MpARF2-DBD in complex with 21ds and ER7 and the structure of AtARF1-DBD in complex with 21ds allowed us to analyse the

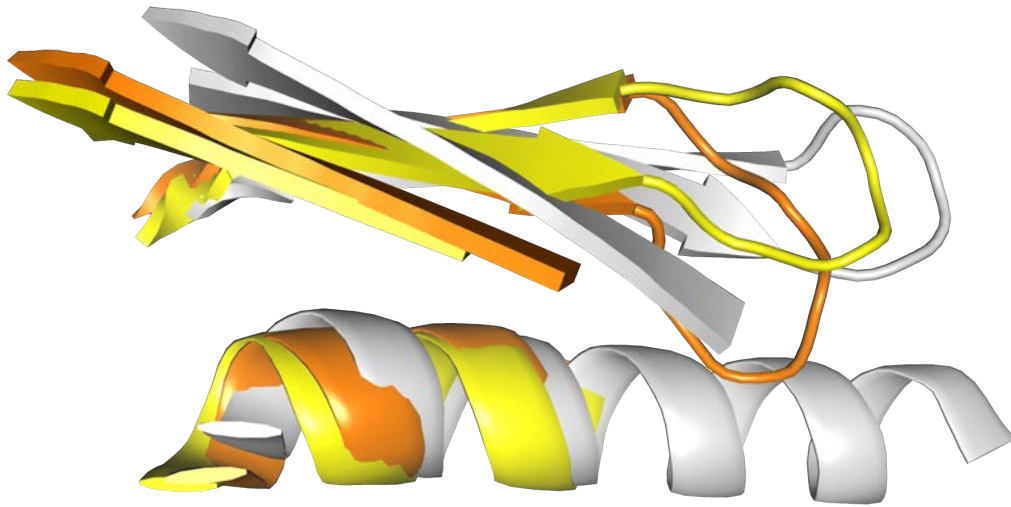


Figure 2.6: B3 β strand length is correlated with $\alpha 1$ length in ARF-DBDs. Longer β strands 4, 9 and 10 and $\alpha 1$ establish more contacts, resulting on higher structuration. At1:21ds is coloured in grey, HR-Mp2:21ds in yellow and ApoAt5 in orange.

structural determinants of the higher affinity of ARFs towards 21ds compared to ER7. In the case of MpARF2-DBD, only the DNA-bound structures with the $I2_12_12_1$ space groups are considered for this comparison, as both display the highest resolution: 2.55 and 2.56 Å for 21ds and ER7, respectively, compared to 2.96 Å of the LR-Mp2:21ds structure.

The contact points of the proteins with the 21ds and the ER7 structures are invariant, except for residues H136 and G137 (AtARF1 numbering, H178 and G179 in MpARF2 numbering). These residues directly contact the GG bases of the high affinity DNA sequence TGTCCG (Figure 2.7) and the last TC bases of the ER7 TGTCTC sequence (Figure 2.8A). The H136 side chain flips over and thereby comes into proximity of the two GG bases, which allows it to make hydrogen bonds to the O6 atom of both, depending on the orientation of the imidazole ring (Figure 2.7B, MpARF2 in complex with 21ds, Figure 2.8B, AtARF1 in complex with 21ds). In the ER7 structure, this orientation of the H136 sidechain is not possible due to steric hindrance of the cytosine N4 atom and the guanine O6 atom of the cytosine-guanine base pair (Figure 2.7C, MpARF2 in complex with ER7, Figure 2.8C, AtARF1 in complex with ER7). The histidine flip is clearly observed in the electron density maps contoured at 1σ , and no other density is observed in other possible sidechain conformers (Figure 2.7D and Figure 2.8D for MpARF2 and AtARF1, respectively). In addition, the loop containing residues H136 and G137 in the 21ds structures is displaced 1.4-1.8 Å in AtARF1-DBD and 1.3-1.8 Å in MpARF2-DBD (residues S134-L142 in MpARF2) compared to the corresponding ER7 structures, due to the extra contact point with G_5/G_6 . This extra contact point also induces a DNA displacement away from the protein in the region encompassing bases $C_4 - G_5 - G_6 - C_7$ by about 1.9 Å in AtARF1 and 0.6-1.2 Å in MpARF2, making room for H136 to penetrate deeper into the major groove. The displacement of the DNA is local, as bases that are further up- and downstream of the G_5G_6 bases coincide with the ER7 structures.

The ability of H136/H176 to change their conformation to accommodate DNA sequence variations could be also important to epigenetic regulation. The regions that show most displacements are localized in the CGG oligonucleotides of the TGTCCG sequence, where the movements in

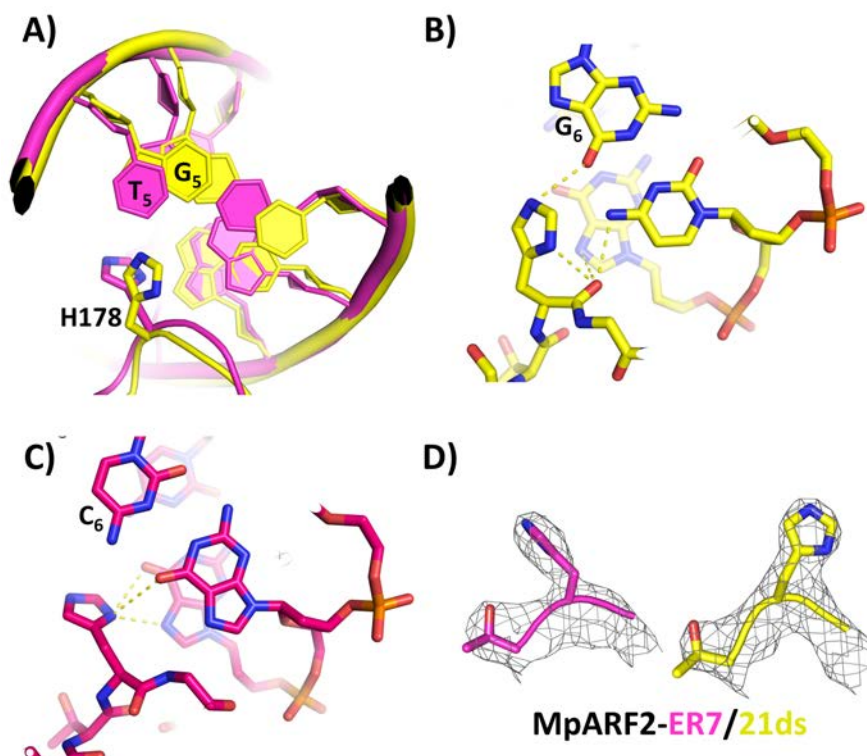


Figure 2.7: Structural details of the MpARF2-DBD:21ds/ER7 complex. A) Overlay of the H178-G179 residues and interacting DNA bases of the ER7 structure (shown in magenta) and the high affinity structure (shown in yellow). It can be clearly seen that the DNA is displaced upwards in order to make room for the entry of H178 into the major groove, allowing it to interact with G_5 of the 21ds sequence. B) and C) Show in detail the interaction between H178-G179 and the G_5G_6 bases in 21ds and T_5C_6 bases in ER sequences, respectively, which are determinant for the affinity of the DNA. D) Detail of the loop containing H178 (T177-G179) showing the 2FoFc electron density map contoured at 1σ , in the ER7 (shown in magenta) and 21ds (shown in yellow) structures.

T135-G137/T177-G179 loops and on the DNA chain contribute to the overall binding of the high affinity 21ds sequence. CGG is a known DNA methylation motif, an epigenetic modification that may affect ARF binding. According to the reported structures, the distances between H136/H176 to the methylable nucleotide atoms in AtARF1 and MpARF2 would be 3.5 and 3.4 Å in the closed conformation (observed in 21ds structures) and 6.1 and 7.4 Å in the open conformation (observed in ER7 structures), respectively, and methylation would have a great impact on the contacts in this region and therefore on the binding affinities. This is in line with episcistrome analysis that suggested AtARF5 as being regulated by DNA methylation [68]. Further biochemical studies are required to ascertain the effect of methylations in AuxRE and the implications of these methylations in ARF-DBD binding.

2.4 Discussion

In this chapter we provided the first X-ray structures of *Marchantia polymorpha* ARF-DBD in complex with promoter-like DNA sequences **ER7** (Mp2:ER7,(I₂I₂I₂I₁)) and the **21ds** high affinity sequence (LR-Mp2:21ds, PDB 6SDG (C2) and HR-Mp2:21ds, (I₂I₂I₂I₁)), an inverted repeat

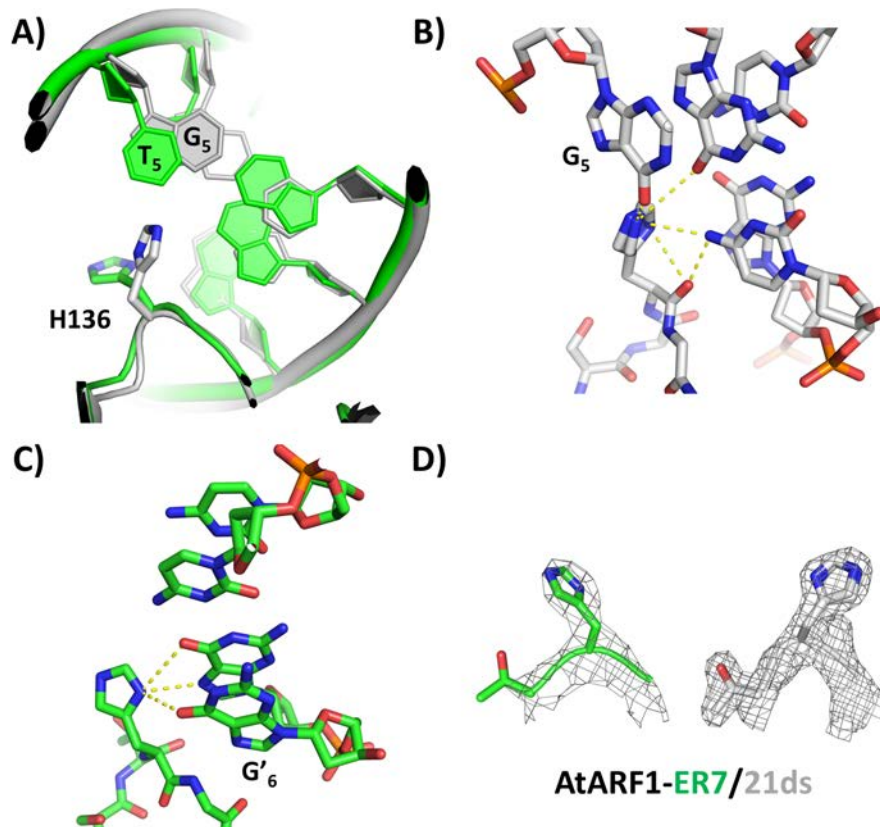


Figure 2.8: Structural details of the AtARF1-DBD:21ds/ER7 complex. A) Overlay of the H136-G137 residues and interacting DNA bases of the ER7 structure (shown in green) and the 21ds structure (shown in grey). It can be clearly seen that the DNA is displaced upwards in order to make room for the entry of H136 into the major groove, allowing it to interact with G5 of the TGTCGG sequence. B) and C) Show in detail the interaction between H136-G137 and the G5G6 bases in High affinity and G6G7 bases of the complementary strand in ER, respectively, which are determinant for the affinity of the DNA. D) Detail of the loop containing H136 (T135-G137) showing the $2F_0F_c$ electron density map contoured at 1σ , in the ER7 (shown in green) and 21ds (shown in grey) structures.

sequence that is based on the TGTCGG recognition element and that was previously found to bind to AtARF1 and AtARF5 with high affinity [39]. We also provide the structure of *Arabidopsis thaliana* ARF1-DBD in complex with this high affinity sequence (At1:21ds, PDB 6YCQ ($P12_1$ 1) [51]). These structures demonstrate the overall structural similarity between all the ARF-DBDs at critical points important for their function.

The structures further deepen our understanding of the molecular calliper model, as they confirm that specific sequence determinants in ARF-DBDs cannot explain distinct affinities for inverted repeats with different spacings, suggesting that certain ARF-DBD structural elements can alter the specificity to accommodate differently spaced AuxREs. In the analysed structures, we have seen that the relative orientation between Dimerization Domain and B3 and the B3 spacing in the dimer are the most variable structural differences in the DBDs. DNA-bound AtARF1 presents a big difference in the relative conformation of B3 and DD compared to the relative position found in ApoAtARF5 and MpARF2 DNA-bound structures. DNA-bound AtARF1 is displaying a DD-B3 planar conformation, where the axis perpendicular to the dimerization interface is nearly parallel

to the DNA axis. In contrast, the dimerization domain axis in AtARF5 and MpARF2 structures is significantly rotated with respect to the DNA axis, with the dimerization domains protruding from the AD part. Without Apo MpARF2 and AtARF5:DNA structures is difficult to determine the degree of change in conformation of MpARF2 and AtARF5 upon binding to DNA. Despite this, we can deduce from the similar difference of the DD rotation between ApoAtARF5 and MpARF2 DNA-bound structures (8°) and the rotation induced upon DNA interaction in the ApoAtARF1-AtARF1 DNA-bound structures (8°) that the planar structure observed in AtARF1:DNA complexes is an intrinsic characteristic of this protein and not induced upon DNA binding. As the Apo state of AtARF1 starts from a higher degree of rotation (169°) than in ApoAtARF5 (157°), the 8° rotation induced in the protein structure by DNA binding would add more torsional forces to the dimerization interface than in the MpARF2 and AtARF5 structures. This sets the AtARF1 DBD in a more restrictive apo conformation, which would impede the binding of oligonucleotides with certain AuxRE spacings, more so than in the case of AtARF5 and MpARF2. This would explain the ability of AtARF5 to interact with a wider range of spacing and suggests a similar behaviour for MpARF2, as both these structures show similar conformations. Experimental analysis should test if this structural similarity between MpARF2 and AtARF5 can be confirmed *in vitro* and *in vivo*.

The DD regions that determine the position of the B3 domain are the $\alpha 1$ helix and the flexible loops connecting B3 and dimerization domain. Judging by the structures, the most important factor on DBD conformation is probably the $\alpha 1$ helix, as the flexible loops connecting both domains are disordered. We have shown that the $\alpha 1$ helix is extended towards the N-terminus in AtARF1, reaching up to 19 amino acids in length. In contrast, MpARF2 and AtARF5 possess a smaller $\alpha 1$ helix compared to AtARF1, of 10 amino acids in length, where the first 9 residues that are helical in AtARF1 are disordered in the structures, despite reasonable sequence conservation with AtARF1. This lack of structure in the N-Terminal part of the helix for MpARF2 and AtARF5 may be produced by the lack of residues near $\alpha 1$ helix that could stabilize this region. It is possible that differences in affinity between both regions explain the $\alpha 1$ helix structuration in AtARF1, as we observed longer β 4, 9 and 10 strands in AtARF1 compared to MpARF2 and AtARF5, which may define the final length of the structured $\alpha 1$ helix. Previous structural studies [39] suggested this helix acts as a pivot point for B3 domain. The new structures suggest that the length of this helix influences the conformation of the B3 domain relative to the dimerization domain, because a longer $\alpha 1$ helix will enable more contacts between Dimerization Domain and B3, thus producing increased separation between the B3s and an overall more planar structure as described above. The increase in separation between the B3 domains and the planar structure in AtARF1-DBD suggests a tendency of AtARF1-DBD to interact with more spaced AuxREs, compared with what is observed in MpARF2 and AtARF5.

In contrast to what would be expected based on the similar classification of MpARF2 and AtARF1, we observe that the structure of MpARF2 in complex with both ER7 and 21ds is more similar to ApoAtARF5, and superposes with lower RMSDs with ApoAtARF1 than with AtARF1 in complex with ER7 and 21ds. MpARF2 structures in complex with DNA are not as planar as AtARF1 upon DNA binding, maintaining an orientation of the dimerization interface similar to that of ApoAtARF5-DBD. This questions the proposed classification based on sequence homology, which groups MpARF2 and AtARF1 on class B, and AtARF5 on class A. The structures indicate that MpARF2 and AtARF5 could prefer smaller spacing than AtARF1, as the overall structure of MpARF2 remains similar to ApoAtARF5 when interacting with the AuxRE spaced by 7 nucleotides, contrasting with the big changes induced in AtARF1 when interacting with a 7 nucleotides AuxRE spacing. This suggests a preference of MpARF2 and AtARF5 proteins to AuxRE separated by a 7 nucleotides, in contrast to AtARF1, which has a preference for larger AuxRE spacings. This is in good agreement with experimental evidences that showed a higher tolerance of AtARF5 to

different spacing compared to AtARF1. Experiments in solution should confirm these observations, to discard the possibility that longer AtARF1 helix is induced by the crystallographic packing.

The structures of the high affinity DNA complexed with MpARF2-DBD and AtARF1-DBD reported in this chapter provide clues on the underlying basis for the increased affinity. The entry of H136 in AtARF1 or the homologous H178 in MpARF2 into the DNA major groove increases the hydrophobic interaction surface of this sidechain with the DNA bases. In addition, it allows for additional hydrogen bonding interaction of distance 2.75 Å between the carbonyl of the H136-G137 amide bond and the N4 of complementary cytosine of G_6 in the TGTCGG sequence. In the ER7 structure, the bulky, complementary guanidine of the C_6 in the TGTCTC sequence, and the methyl group of T_5 likely prevent the H136 from approaching these bases, preventing the formation of the favourable hydrogen bond and hydrophobic interactions observed in the high affinity structure. It must be taken into consideration that all members of class C have a substitution of the histidine in this position to a glycine. This substitution could prevent Class C ARFs to interact with the DNA with a high affinity as found in class A and B, highlighting the divergence between classes A/B and C experimented during evolution.

An unexplored field in ARF gene regulation is the possible effect of DNA methylation on ARF-DBD binding. Plant genomic DNA methylation is a very common event that happens in 1 of each 7 cytosine bases [63]. The explanation of this high rate of cytosine methylation is that it occurs independently of sequence, although methylation on CG sites is the most common, as is the case in vertebrates. This fact implies that the ER7, TMO5 and 21ds sequences can be affected by cytosine methylation. Indeed, the structures of ARF-DBD:DNA shows that the position of 5-methyl group of a putative 5-meC would be at a suitable distance to DBD amino acid sidechains to have an effect on DBD:DNA interactions [69, 70]. This suggests that a methylation at C_4 of the auxRE could alter the interaction with H136, which is involved in the recognition of 21ds. The implications of these observations should be addressed in future studies due to the possible implication in ARF gene regulation.



3. Testing the calliper model in solution

3.1 Abstract

The first ARF2-DBD crystallographic structures of the *Arabidopsis* distant relative *Marchantia polymorpha* showed that MpARF2 is able to interact with ER7 and 21-7 dsDNAs using a similar mechanism to that found for AtARF1. To get insights into the calliper model of ARF-DNA interaction, we analysed the binding of MpARFs and AtARFs to a variety of DNA sequences, by means of Analytical Size Exclusion Chromatography, Small-Angle X-Ray Scattering and Fluorescence Anisotropy assays, which revealed that all ARFs are able to dimerize, although the presence of a suitable spaced palindromic AuxRE is the main driving force for dimerization. Our results show variable affinity for DNA among ARF classes, where the interaction profile of class C MpARF3 differs from that of class A and B ARFs. We also tested the ability of ARFs to heterodimerize. Our combined results suggests that, as proposed by the molecular calliper model, ARFs bind as a dimer to auxREs separated by seven or eight nucleotides, and as a monomer to all other.

3.2 Introduction

The molecular calliper hypothesis was proposed based on the first ARF-DBD X-Ray structures: different ARF will vary in the flexibility of their DBD subdomains, giving tolerance and restraints to bind oligonucleotides with differentially spaced inverted repeats of AuxREs. Knowing the spacing tolerance of ARFs will allow us to predict the promoters under control of each ARF. This model leaves unanswered questions on the contribution of the high-affinity calliper binding to the total cellular auxin response. Some studies consider that this calliper model cannot explain the ARF binding to all auxin responsive promoters, as the presence of inverted repeat arrangements is not as common as single AuxREs in the promoters of auxin-response genes [27]. Taking this into account, our aim was to further refine the model based on *in vitro* information of protein-

DNA and protein-protein interactions to better understand ARF signalling. For this purpose, we analysed several AuxRE-like DNA sequences from promoters of genes under ARF control (Figure 3.1), including **TMO3** (GGTCAAaagtaagacTGGACC), **TMO5** (GGTCTCtggtcggTCGACA) and **LFY** (TGTCAAAtttcccagcAAGACA) [39]. In the case of the promoter of TMO5, a gene under AtARF5 control, we included two more sequences with a single (**TMO5 Δ 1**) or complete deletion of the AuxREs (**TMO5 Δ 2**). We also analysed the high affinity **21ds** sequence (TGTCGG) with variable AuxRE spacing ranging from 5 to 9 nucleotides and the canonical ER (TGTCTC) with spacing of 7 and 8 nucleotides to investigate the impact of variable spacing on the binding of ARFs from different classes. Finally, we included the 21ds sequence in direct repeat configuration with spacing of 5 nucleotides (**DR5**), with the aim to study if different AuxRE arrangements not covered by the molecular calliper model can interact with ARF-DBDs. The interaction of an ARF dimer with two consecutive direct repeats would require a separation of 5 nucleotides between the auxREs.

AuxRE1 - 2 - 4 - 6 - 8 - AuxRE2

ER7 : TGTCTCCTTTGG--GAGACA
ER8 : TGTCTCCAAAAGG-GAGACA
21-5 : TGTCGGCATTG----CCGACA
21-6 : TGTCGGCGATCG---CCGACA
21ds : TGTCGGCGATTTCG--CCGACA
21-8 : TGTCGGCGATATCG-CCGACA
21-9 : TGTCGGCGATTATCGCCGACA
LFY : TGTCAAATTTCCCAGCAAGACA
DR5 : TGTCTCCTTT-----TGTCTC
TMO3 : GGTCAAAGTAAGACTGGACC
TMO5 : GGTCTCTGGTCGG--TCGACA
TMO5 Δ 1 : GGTCTCTGGTCGG--TTTTTT
TMO5 Δ 2 : TTTTTTTGGTCGG--TTTTTT

Figure 3.1: Rationale behind DNA sequence design for *in vitro* testing. The AuxRE-like sequences are coloured in red, while spacers are shown in black.

The new structures presented in chapter 2 show that *Marchantia polymorpha* ARF-DBD, like in *Arabidopsis thaliana*, interact with inverted repeats separated by seven nucleotide as a dimer, suggesting a conserved mechanism of DNA recognition in highly diverged plant species. ARF dimerization is expected to contribute to transcriptional regulation because dimer formation prior to interaction with palindromic AuxREs can increase affinity for the binding site due to cooperativity. The formation of oligomers can represent an additional regulatory element, where protein concentration may determine the chances of homo- or heterooligomerization, or variable relative concentrations of different ARFs in the cell may influence the type of heterooligomer formed [39, 71]. Similarly, heterooligomerization could alter the DNA binding specificity of the partners recruited by the transcription factor, leading to complex cellular responses to changes in the cellular context [71]. Despite the possibility of ARF heterooligomerization, few studies have demonstrated that it occurs *in vitro* [32]. To date, studies have shown heterooligomerization to occur *in vitro* between different full length ARFs [29], ARF C-Ter PB1 domain [72] or ARF-DBD in complex with DNA [45]. The PB1 domain is essential for ARF heterotypic interaction, mainly with other PB1 domain-containing proteins such as Aux/IAA proteins. PB1 domains are a common module for protein-protein interaction, so the PB1 domain of different ARFs could contribute to homo- or even heterooligomerization. Studies in *M. polymorpha* plants revealed that full length

MpARF1 and 2 are able to heterodimerize [73], indicating a strong relevance of ARFs coordinated action. Some ARFs with a truncation in PB1 domain lose their ability to form ARF-Aux/IAA or ARF-ARF complexes, supporting the implication of PB1 in ARF homo- and heterooligomerization [74]. In contrast, deletion of this domain in the ARF5 and ARF7 activators only reduced auxin responsiveness by a fraction, highlighting that these two ARFs do not require their PB1 domain for auxin responsiveness [27, 75]. While those studies showed that ARF heterooligomerization occurs for some ARFs, it is still not clear whether heterooligomerization is driven only by DNA binding, by C-Terminal protein-protein interaction PB1 domain or if formation of heterodimers is an intrinsic ability of the Apo ARF-DBD [32].

In this chapter we test the previously proposed calliper model of DNA interaction by means of Analytical Size Exclusion Chromatography, Small-Angle X-Ray Scattering and Fluorescence Anisotropy assays. Furthermore, *in vitro* tests with native ARFs in absence of DNA were performed in order to clarify the contribution of the DBD in ARF dimerization. We analysed the behaviour of 4 different members of the ARF family from *Arabidopsis thaliana* and *Marchantia polymorpha*, representing the 3 classes. This allowed us to compare the behaviour of evolutionary distant homologues.

3.3 Results

3.3.1 ARFs show sequence and spacing dependent affinity *in vitro*

We studied the ARF-DBD DNA binding interactions by analytical Size Exclusion Chromatography (SEC) and Fluorescence Anisotropy (FA), using a set of different recognition sequences in **inverted** (TMO3, TMO5, LFY, 21, ER), and **direct** (DR5) **repeats**. Furthermore, variants of **21ds** with variable spacing (from 5 to 9 nucleotides) and of **ER** (7 and 8) sequences were analysed for contribution on protein binding affinity (Figure 3.1). Mixtures of class A or B ARF-DBDs with double stranded oligonucleotides containing inverted repeats of ER and 21ds AuxREs spaced by 7 or 8 nucleotides show shifts in SEC profiles, resulting in elution volumes compatible with ARF-DBD homodimers in complex with the dsDNA, according to our in-house calibration. This suggests that the tested ARFs can accommodate inverted AuxRE spacings of 7 and 8 base pairs. When spacings differ from the optimal 7-8 nucleotide, interactions are still observed for MpARF2, as shown by peak shifts in SEC chromatogram. However, the elution volume of this species does not correspond to the homodimer:DNA complex. In this case, partial dimer:dsDNA complex formation can explain the lower peak mobility compared to 7-8 nucleotides spacing, for example, or a complex formed by an ARF monomer and the dsDNA may also be an alternative, indistinguishable explanation of the results (Figure 3.2).

Contrary to AtARF1, which do not show a peak shift with a nine nucleotide spacing, AtARF5 is showing partial shifts in the case of a spacing of 9 nucleotides, and both AtARF1 and AtARF5 show no interaction with 5-6 nucleotides of spacing, probably due to a lower affinity for oligonucleotides with a spacing smaller than 7 nucleotides. The results also show that under the tested conditions, none of the tested ARFs showed binding to the 21ds sequence in direct repeat configuration (Figure 3.2). Interestingly, neither LFY nor TMO3, two *in vivo* target sequences with 9 nucleotides of spacing, interact with any of the tested ARFs (AtARF1, AtARF5, MpARF2 and MpARF3) under the tested experimental conditions, which suggest that other factors could be required to help ARF interact with these sequences *in vivo*. In the case of TMO5, this *in vivo* sequence is a hybrid of ER and 21ds sequences with a mutation at each end (GGTCTC vs TGTCTC (ER7) and TGTCGA vs TGTCGG

(21ds)). Only MpARF2 and AtARF5 were tolerant to these changes and displayed shifts on SEC assays with the TMO5 sequence (Figure 3.2).

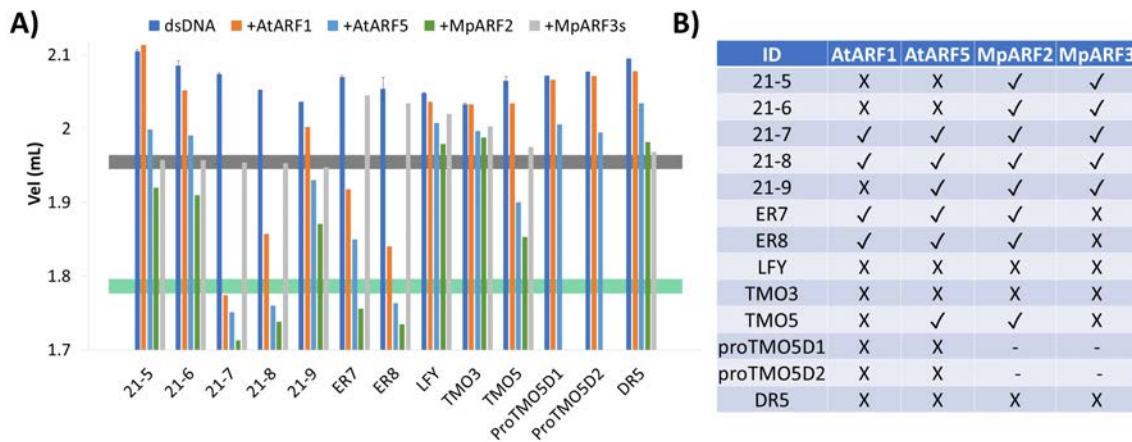


Figure 3.2: Size exclusion peak shift analysis of AtARF1, AtARF5, MpARF2 and MpARF3 DBDs in complex with several AuxRE-like DNAs. A) Elution volume corresponding to each dsDNA (dark blue) or dsDNA in complex with AtARF1 (green), AtARF5 (light blue), MpARF2 (green) or MpARF3 (grey) DBDs at a constant dsDNA:ARF-DBD molar ratio of 1:2. Horizontal black- and green-shaded horizontal bars represent the theoretical elution volume calculated for a dsDNA in complex with an ARF-DBD monomer or dimer, respectively. B) Qualitative analysis of the results. An interaction was assigned when the elution volume was lower or equal to the elution volume corresponding to a monomer:dsDNA.

SEC analyses showed that the interaction with AuxREs is spacing and sequence dependent, where an optimal spacing of 7-8 nucleotides was found, but partial shifts with other spacings are also observed. The interpretation of partial shifts is ambiguous and do not give information about the affinity or the binding mode to the DNA. Since SEC only provides qualitative data, fluorescence anisotropy assays were performed to determine quantitatively the affinity of ARFs towards AuxRE sequence and spacing.

The fluorescence anisotropy results show that AtARF1 and AtARF5 have higher affinity towards 21 AuxRE than to ER, TMO3, TMO5 and LFY (Table 3.1). In addition, the optimal spacing is found to be 7 or 8 nucleotides. Saturation curves for spacing different than 7-8 follow a Michaelis-Menten mechanism, whereas a more complex interaction is found in curves on 7-8 spacing. In this latter case, an initial rapid increase of the anisotropy at low protein concentrations suggests that a quick interaction with DNA is initially produced. At higher protein concentrations, the increase is slower (see Supplementary information Figure 9.2). Overall, this suggests a first step of interaction with one AuxRE with high affinity, followed by a second step in which a structural rearrangement is produced to accommodate for the interaction with the second AuxRE. When the spacing is not compatible with binding of an ARF dimer, the second step is not produced, and the saturation curve shows the interaction with only one of the AuxREs. We can also see that AtARF1 preferred the interaction with 21-8 rather than 21-7, while AtARF5 has higher affinity towards 21-7 than 21-8. All the other interactions with non 7/8 spacing were stronger for AtARF1, but the model fits of AuxRE spacings different from 7 and 8 are difficult to interpret. The best fits suggest an interaction as monomer in AtARF1 and dimer in AtARF5, which may be decreasing the apparent AtARF5 affinity towards short-spaced AuxRE, as the structural rearrangement required to bind those spacing is big. FA results also show LFY interaction in FA with AtARF5 but not with AtARF1, partially agreeing with SEC results. In any case, as the fits with AtARF5 resulted in high errors, we can not

Table 3.1: ARF:DNA interaction parameters obtained from fluorescent anisotropy

| DNA | AtARF1-DBD | | AtARF5-DBD | |
|------|----------------------|------------------|----------------------|------------------|
| | Kd (μM) | Hill coefficient | Kd (μM) | Hill coefficient |
| 21-5 | 4.18 \pm 1.36 | 1 | 26.29 \pm 10.18 | 0.56 \pm 0.05 |
| 21-6 | 3.24 \pm 0.78 | 1 | 18.02 \pm 8.92 | 0.64 \pm 0.09 |
| 21-7 | 2.87 \pm 0.063 | 0.37 \pm 0.03 | 0.90 \pm 0.22 | 1 |
| 21-8 | 0.81 \pm 0.20 | 0.47 \pm 0.05 | 2.33 \pm 0.44 | 0.90 \pm 0.22 |
| 21-9 | 3.18 \pm 1.00 | 1 | 8.79 \pm 1.22 | 0.77 \pm 0.05 |
| ER7 | 28.06 \pm 7.49 | 1 | 3.74 \pm 0.96 | 1 |
| ER8 | 15.86 \pm 4.06 | 1 | 0.43 \pm 0.31 | 1 |
| DR5 | 9.88 \pm 2.67 | 1 | 22.97 \pm 7.56 | 0.97 \pm 0.13 |
| TMO5 | 0.71 \pm 0.10 | 0.58 \pm 0.07 | 6.52 \pm 1.63 | 0.73 \pm 0.11 |
| LFY | NI | - | 9.03 \pm 2.49 | 1.18 \pm 0.27 |

Hill coefficient of 1 was assigned when the best fit was obtained with a Michaelis-Menten fit

NI: No Interaction observed

conclude from these results that AtARF1 and AtARF5 have a different spacing preference. The high errors observed in AtARF5 may be explained by its low stability in the assay buffer, with low sodium chloride concentration. This is the reason why we were not able to perform the fluorescence polarization measurements with MpARF2, as due to its instability at low salt concentrations made it impossible to reach the high concentrations required to perform the assay.

3.3.2 DNA binding drives ARF dimerization

To address the contribution of protein concentration to dimerization, Small-Angle X-ray Scattering (SAXS) was performed on a range of concentrations of different Apo-ARF-DBDs. In addition, in order to elucidate the contribution of DNA binding and the effect of spacing variants of the high affinity DNA sequence on ARF dimerization, DNA:ARF-DBD complexes at variable stoichiometry were analysed by SAXS. Analysis of ApoARF-DBD SAXS profiles using the program OLIGOMER [76] suggests that all analysed class A/B ARFs are able to homodimerize in solution under the tested conditions. In all class A/B ARFs, the particle size increased with protein concentration (see supplementary Table 9.7). Neither monomer nor dimer models alone explain the experimental data. Results show that models of equilibrium between dimer and monomer improved the fit dramatically, and that the distribution over dimeric and monomeric forms is dependent on protein concentration (Figure 3.3).

It can also be appreciated that AtARF5-DBD and MpARF2-DBD behave similarly, as both are almost exclusively in the dimeric form at a concentration of 3.5mg/ml. In contrast, 60% of AtARF1 is in a dimeric state at 3.5mg/ml (Figure 3.4). It is also relevant to highlight that at higher concentrations than 1.75mg/ml in all the tested ARF-DBDs the Chi^2 of the fits, using a model consisting in a monomer:dimer based on the conformation observed in crystallography, were unacceptable (See supplementary Table 9.5), probably indicating that less stable oligomeric species could be forming in the solution with different quaternary structure than that observed in the crystal structures. In fact, *ab initio* models of the dimers in solution suggest certain variability in the dimerization interface in ARFs that could explain the increase in Chi^2 values, as these models differ substantially from the known crystallographic models used for computations (Figure 9.1). In contrast, AtARF1 requires up to 7mg/ml before a dimer fraction of 80% is reached, indicating that the dimerization K_D could be higher in AtARF1-DBD compared to AtARF5-DBD/MpARF2-DBD.

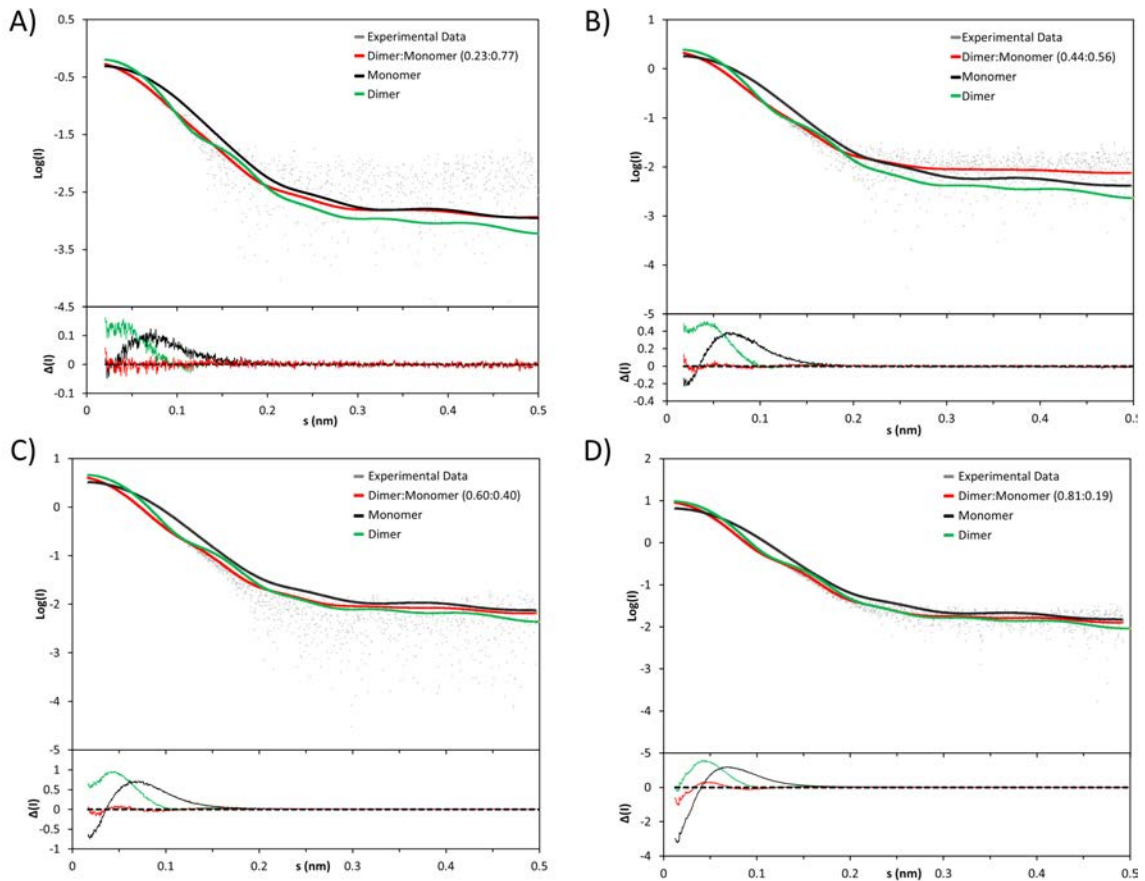


Figure 3.3: Small-angle X-ray scattering curves of solutions containing ApoAtARF1. Top panels: Experimental data points of Apo AtARF1 at 0.7mg/ml (A), 1.75mg/ml (B), 3.5mg/ml (C) and 7mg/ml (D) are shown as grey dots. The dimer fit is shown in green, the monomer fit in black and the fits for Monomer:Dimer mixtures are shown in red. The specific ratio of Dimer:Monomer is specified in figure legends. The bottom panel shows the intensity differences between the fitted curves and the experimental data.

All the MpARF3 concentrations tested resulted in good fits with low χ^2 values for a model consisting in an MpARF3 monomer, using an *in silico* model created using MpARF2 structure as template. This result is in line with the prediction that in MpARF3, the dimerization interface is disrupted by an 80 amino acid insertion in $\alpha 6$ helix, thereby preventing homodimerization [55]. This highlights an important difference of ARF class C in comparison with class A and B. The computation of a chain-like *ab initio* model from SAXS Data using GASBOR [77] also confirms the shape of MpARF3 as a monomer, very similar to the *in silico* model (Figure 3.5A). The only mismatch between these two models is in the 80 amino acids that are inserted into the dimerizing $\alpha 6$ – helix, but which is not found in the MpARF3 SAXS model. This could be explained by assuming that this loop is very flexible and therefore does not give contrast in this area. The SAXS model is more compact in the part connecting the Dimerization domain and B3 compared to the other assayed ARFs, so these 80 amino acids could be also packed into this cavity. The dimeric models obtained from the SAXS profiles of the other ApoARFs are not consistent with the crystallographic dimers, showing alternative dimerization interfaces and positions (see supplementary information, Figure 9.1). The *ab initio* models for apo proteins seem to differ from the crystallographic structures, suggesting that the dimers observed in crystallographic structures

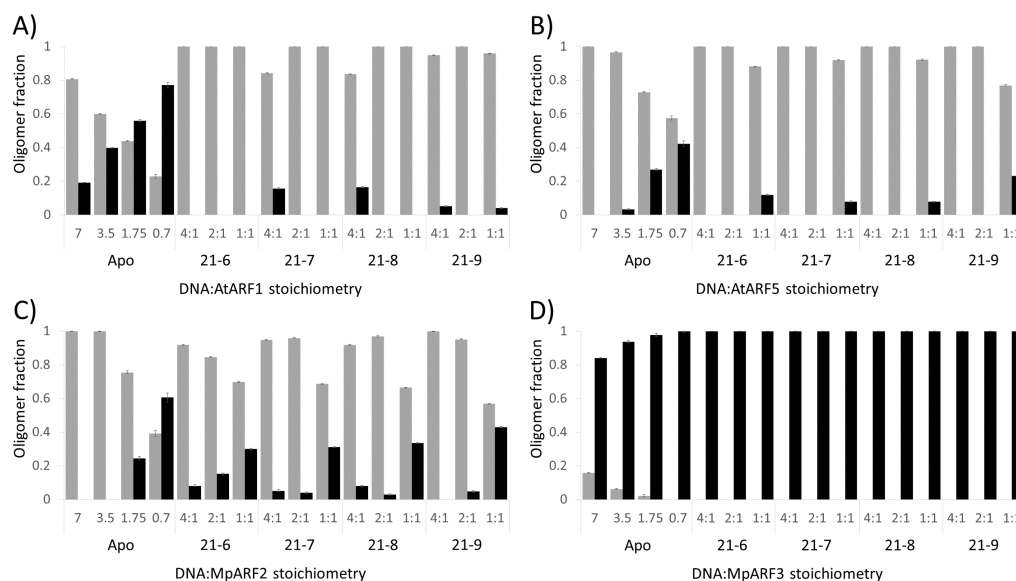


Figure 3.4: Distribution of monomer (black bars) and dimers (grey bars) on ApoARF and DNA:ARF complexes calculated from SAXS data. All the ARF:DNA measurements were done at $[ARF] = 3.5\text{mg/ml}$. A) AtARF1-DBD, B) AtARF5-DBD, C) MpARF2-DBD, D) MpARF3-DBD. The proportion of monomer and dimer were calculated using OLIGOMER.

are induced or favoured by the binding to the DNA. Overall, SAXS data with Apo ARFs highlights the differences between class C and A/B and the high variability in dimer formation.

The SAXS data show that class A and B ARFs are able to dimerize *in vitro* in a concentration dependent manner in the absence of DNA. Next, we addressed by SAXS the contribution of DNA to dimer formation. As for the Apo data, we calculated monomer/dimer distributions of protein:DNA complexes using the program OLIGOMER. It should be noted, however, that the χ^2 values of the fits were quite high compared to those of the Apo proteins, which may be due to the presence in solution of less stable oligomeric species, which are not taken into account in the calculated fits using the crystallographic dimer (see supplementary Table 9.6). The addition of DNA to ARF samples at 3.5mg/ml resulted in the displacement of the equilibrium towards dimer formation, except in the case of MpARF3 (Figure 3.4). The clearest results were obtained with AtARF1 samples, where the addition of DNAs shifted the dimer ratio from 60% to almost 100% in all cases. Interestingly, at a ratio of 4:1 (DNA:Protein) of 21ds, 21-8 and in a lower extent in 21-9, the AtARF1 dimer formation seems hindered and a small portion of monomeric AtARF1 fit the data better (Figure 3.4A). This suggests that in an excess of DNA ligand (AuxREs), AtARF1 could preferentially bind to a single site as monomer rather than to an inverted repeat as a dimer. This is in line with the high AtARF1:DNA-binding affinity and with AtARF1 homodimerization being less favourable compared to AtARF5 and MpARF2, as shown in SAXS results with Apo proteins (Figure 3.4). The effect of DNA addition on the monomer/dimer distribution of AtARF5 and MpARF2 is not as clear as it was in AtARF1 since these proteins in apo form were already found as dimers at concentrations as low as 3.5mg/ml . Even with the high dimerization of MpARF2 and AtARF5 at 3.5mg/ml , an increase in dimer ratio with an increase of DNA:protein stoichiometry is observed for both MpARF2 and AtARF5 suggesting that DNA can drive dimerization of class A and B proteins tested (Figure 3.4B and C). In the case of MpARF3 data, the best fit was obtained for a DNA-bound monomer over all DNA concentrations (Figure 3.4D). Furthermore, the computed model of MpARF3 in complex with 21-6 is the only model with good shape, which shows an

MpARF3 monomer bound to 21-6 dsDNA (Figure 3.5B).

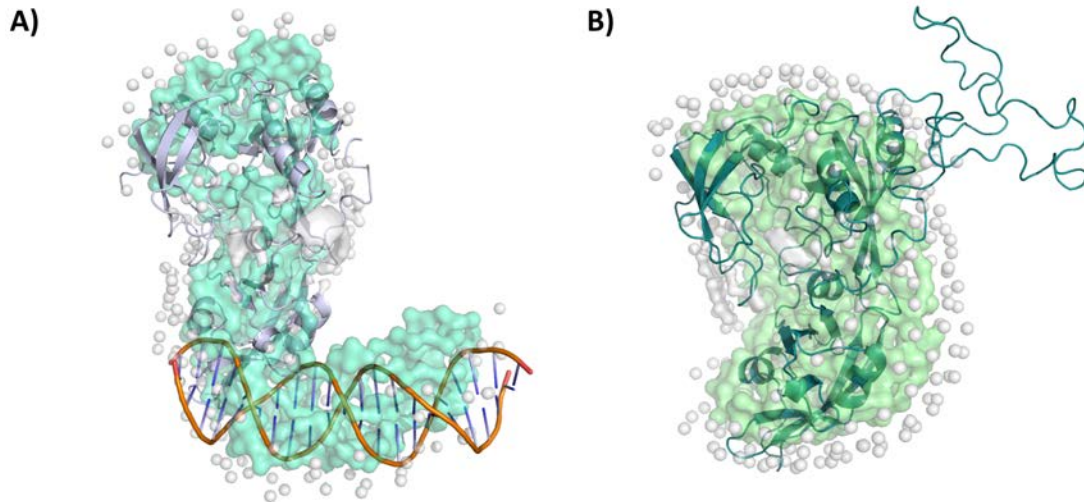


Figure 3.5: *In silico* model of MpARF3. ApoMpARF3 (Left panel) and MpARF3:21-6 (right panel) models were computed with Swiss-Model based on the structure of MpARF2 in complex with 21ds (6SDG). The computed models (cartoon representation) agree with the volumes calculated with SAXS data (ball representation)

To double check the results obtained using fitting of the structural models and *ab initio* reconstructions, we also analysed typical SAXS particle size parameters, R_g , V_{porod} and D_{max} (Table 9.7). These parameters are in good agreement with the calculations on monomer/dimer distribution already presented, where the increase in DNA:protein ratio result in higher values of R_g , V_{porod} and D_{max} for MpARF2 and AtARF5, consistent with a shift of the protein monomers:dimer distribution towards dimers, producing particles with bigger volumes. In the case of AtARF1, all size indicators increase as DNA:protein ratio increases, with the exception of the V_{porod} , which systematically decreases at ratio 4:1. A constant R_g and D_{max} combined with a concomitant reduction of the porod volume indicates an elongation of the overall shape of the complex. This indicates that an AtARF1 monomer bound to DNA, which has an elongated rod-like shape, fits the SAXS curve better than the DNA:dimer complex, which has the shape of a flat cylinder. Finally, in the case of MpARF3, results are consistent with a monomer bound to the DNA, as the parameters vary little when increasing DNA:protein ratio. These results also confirm the MpARF3 monomeric interaction with 21ds with spacings in the 6-9 range observed in SEC analyses, as in all the cases the R_g , V_{porod} and D_{max} parameters of DNA:MpARF3 samples are higher than for Apo MpARF3. We find that the resulting particle in the MpARF3:DNA mixture is bigger than the components alone, but is not as big as a dimer bound to a dsDNA.

3.3.3 The ARF-DBD alone is enough to promote ARF-ARF heterodimerization

ARF heterodimerization can increase the complexity of the auxin response depending on the NAR pathway. As in other transcription factors like animal retinoid receptors that form heterodimers with a huge variety of receptors [78], the ability of ARFs to form heterodimers would help to explain the variability of responses to the simple auxin molecule. ARFs have been proposed to form heterooligomers through their C-terminal PB1 domain, which is a widely known protein:protein interaction module. Also, some authors proposed that DNA binding could allow ARF heterodimerization. We wanted to know if the DNA Binding Domain alone can form heterodimers through

the dimerization domain. We also wanted to know if this dimerization interface was sufficiently conserved during evolution to allow ARFs from different species to heterodimerize. Our approach was based on the Dot-blot method [79]. This method allows analysing the interaction of proteins in native conditions. We spotted drops of different *Arabidopsis thaliana* and *Marchantia polymorpha* proteins in a range of concentrations to estimate an interaction affinity.

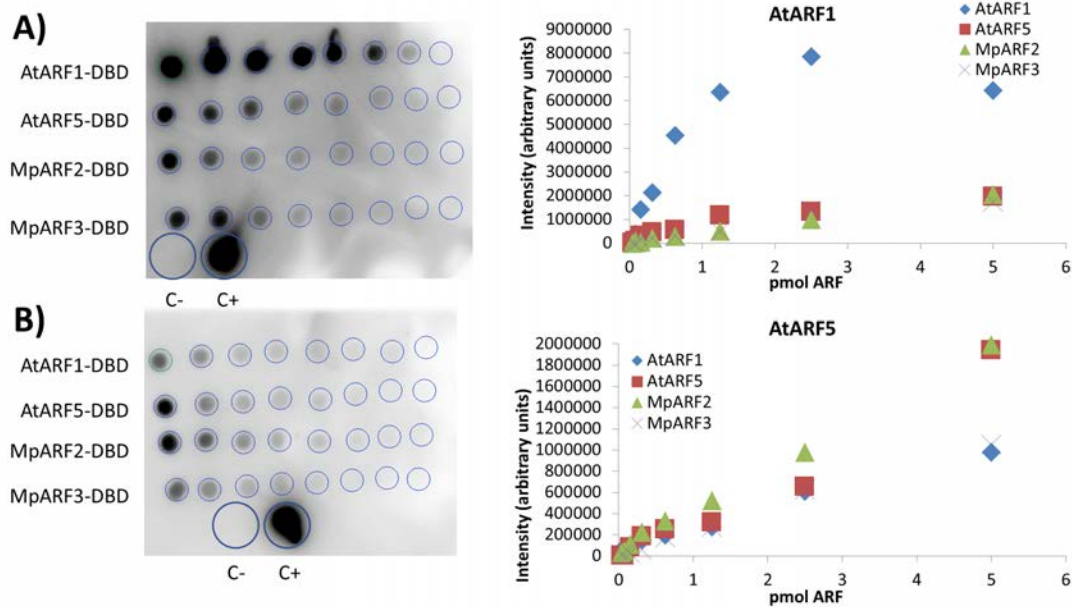


Figure 3.6: Dot-blot assays of spotted ARFs. ARF-DBDs were spotted in two identical membranes in a range of 5-0.04pmol/spot. A membrane was incubated with His-tagged AtARF1-DBD, while B membrane was incubated with His-tagged AtARF5-DBD. The quantification of the spot intensity is presented in the graphs on the right.

Our results show that AtARF1 and AtARF5 can form homo- and heterotetramers, and that they can interact with *Marchantia polymorpha* MpARF2 and MpARF3 (Figure 3.6). In addition, AtARF1 is more inclined to self-dimerize than heterodimerize, as it can be clearly seen from the interaction analysis (Figure 3.6A). AtARF5 is more prone to form heterocomplexes, as similar interaction curves with all the tested proteins are achieved (Figure 3.6B). We would like to highlight that AtARF1 blots show overall higher intensity, as AtARF5 is intrinsically more unstable in solution during long incubations and low salt concentrations than AtARF1. This could lead to lower effective concentrations of AtARF5, resulting in lower intensity.

Interestingly, the dot blot results show that heterodimerization of AtARF1 and AtARF5 with MpARF3 occurs (Figure 3.6). This is surprising because MpARF3 contains an 80 amino acid insertion believed to be unstructured, disrupting the $\alpha 6$ helix that is critical for ARF dimerization. This is confirmed by lack for formation of MpARF3 dimers using SAXS and SEC. However, here we see that this insertion is not preventing heterodimerization. There are three possible explanations. Firstly, the interaction may still be possible despite this insertion, although with low affinity. The low affinity could prevent MpARF3 homodimerization, as observed in SAXS, but other ARFs with an intact dimerization interface could partially counteract, allowing heterodimerization. Secondly, there is the possibility that other low affinity dimerization interfaces, not seen in crystallographic structures, are present in ARF-DBD, allowing MpARF3 to dimerize. Finally, dot-blot may have higher sensitivity and detect MpARF3 heterodimerization, not detected in SEC or SAXS. Further studies are needed to clarify the role of the $\alpha 6$ helix insertion in MpARF3.

3.4 Discussion

The results presented in this chapter deepen our understanding of the functional mechanism of the molecular calliper model. We have tested the ability of different ARFs to accommodate variable AuxREs and spacings, where we observed high similarities between interaction patterns of AtARF5 and MpARF2, despite the fact that they belong to different classes (A and B, respectively). The similar conformation of B3 domain with respect to the dimerization interface observed in crystallographic structures of MpARF2 and AtARF5 may explain the broader tolerance to spacing and recognition sequence of MpARF2 and AtARF5 compared to AtARF1 (see chapter 2). This also agrees with the results of SEC on the interactions with different oligonucleotides, with fluorescence anisotropy experiments and with the model of ARF competition for the binding site previously reported [28]. We also observed that AtARF1 interacts with a narrow range of allowed auxREs spacings. This is in line with the results presented in the previous chapter, where analysis of the crystal structures showed that the B3 and dimerization subdomains of AtARF5 and MpARF2 were less planar, mainly due to a shorter $\alpha 1$ compared to AtARF1 $\alpha 1$. The SEC results described in this chapter with the high affinity 21ds sequence also agree with previous SPR results on interactions between AtARF1 and AtARF5 with ER5-9 DNA sequences. Thus, AtARF1 and AtARF5 bind with similar efficiency to 21-7 and 21-8, whereas AtARF1 has much lower affinity for 21-5, 21-6 and 21-9. In contrast, AtARF5 has lower but similar affinities to 21-5, 21-6 and 21-9 compared to 21-7 and 21-8 [39]. However, the smaller peak shifts of AtARF5 in complex with 21-5, 21-6 and 21-9 oligonucleotides suggest that AtARF5 is bound to the DNA for a smaller fraction of time. In a cellular context, this could mean lower expression activation as the proteins spends less time bound to the promoter, which would result in fine tuning of the expression response with AtARF5 binding.

SEC and SAXS results are also in line with previous results that show that Class C ARFs behave differently from A/B ARF classes. Our results show class C ARFs interact with AuxRE exclusively as monomers, and the Apo and DNA-bound MpARF3 is always found as monomer in solution. This is not the case of A/B ARF classes, which show a tendency to dimerize *in vitro* in a concentration-dependent manner, where DNA drives the formation of dimers. We cannot discard that dimers can form in manner different from that observed in crystallography, as SAXS *ab initio* models proposed that Class A and B could form different types of dimers in solution in absence and in presence of DNA. The dimer observed in crystallography only fits with SAXS data in the presence of 21-7 or 21-8 oligonucleotides. When oligonucleotides with different spacings are present, the protein also seems to dimerize, but through different dimerization interfaces. These different dimerization modes could allow ARFs to interact with everted and direct repeats, two mechanisms of binding that cannot be explained with the dimerization interface observed in crystallography.

SAXS analysis of ARF solutions at different concentrations also highlights the similarity between MpARF2 and AtARF5, comparable to the structural relationship that was already observed in the crystal structures presented in chapter 2 and in SEC experiments described above. MpARF2 and AtARF5 also show similar DBD heterodimerization behaviour, where again AtARF5 is more permissive in the dimerization to other ARFs than AtARF1. In fact, AtARF1 preferentially forms homodimers, although all the tested proteins were able to interact with AtARF1. In contrast, AtARF5 does not show evident preference for any ARF, and is able to heterodimerize with many, even with distant homologues. Despite belonging to different classes, the DBD of these two proteins are very similar in their structure and *in vitro* properties. Both share similar K_D s in SAXS, similar interaction pattern in SEC and similar main chain position in both the ER7 and 21ds-bound crystal structures. As DBDs are very conserved and most sequence variations in ARFs are concentrated in the MR, the classification will be strongly biased by the MR sequence. This suggests that

ARF classification based on full length sequence are not considering differences on the different domains, and explain why MpARF2, classified as class B, seems to contain a DBD similar to class A AtARF5. The combination of DBDs from different classes may explain why complex organisms as *Arabidopsis thaliana* encode for many ARF homologs. Probably, the combination of domains from different groups has arisen due to an evolutionary advantage by an increase of ARF variability, expanding complexity and allowing more precise gene regulation. Taking this into account, we believe that ARF classification may consider ARF independent domains to understand the variability inside each ARF class. This classification could include key observations, such as the $\alpha 1$ – helix length and DNA binding preference, apart from the sequence homology.

Remarkably, we found heterodimerization of AtARF1 and AtARF5 with MpARF3, a class C ARF, considering that SAXS indicates that apo MpARF3 is unable to homodimerize. Thus, there is the possibility that class A and B ARFs may be able to interact with class C ARFs because they have a properly formed dimerization interface. Thus, the insertion found in helix $\alpha 6$ of MpARF3 would prevent homodimerization, but could still allow forming contacts with other ARFs which have an undisturbed dimerization interface. This is in line with the competition model of ARF gene selection²⁸, where dimerization and gene transcription are dependent on the relative concentrations of ARFs from different classes and on AuxRE availability. Regulation of the expression of genes controlled by ARFs can be fine-tuned by controlling the concentration of ARFs through expression, degradation, translocation or interaction with other partners and/or the relative availability of ARFs from different classes.

In conclusion, our results suggest that the relative positions of the subdomains of the DBDs determine the specificity of the ARFs to the promoter sequence, as proposed by the molecular calliper model. We have seen that the structural similarity between AtARF5 and MpARF2 reported in chapter 2 is producing a similar DNA binding specificity *in vitro*. This give information on the most important structural elements in determining the affinity towards differentially spaced AuxREs. *In vitro* results obtained in this chapter can be explained by the ability of ARFs to increase the distance between B3 domains of the dimer, which is required to accommodate larger AuxRE spacings, as described in chapter 2. Due to helical nature of the DNA structure, as spacing increases the position of the major groove rotates along the DNA axis, requiring a rotatory adaptation of the ARF structural changes to interact with the AuxRE, apart from an adaptation of the distance in B3 domains. As structures showed, the B3s of AtARF1-DBD are more separated and the B3/DD twist angle is smaller than in AtARF5 and MpARF2, which may help to accommodate AuxREs with larger spacings.

The aforementioned mechanism to create specificity may not be enough as there are limited combinations of AuxRE spacing to allow ARF interaction as dimer. In complex organisms as *Arabidopsis thaliana*, where 23 ARFs exist, other elements of the ARF structure may determine the binding affinity. Protein-protein interaction domains that recruit cofactors that localize ARFs on the binding site is a possible element to modulate affinity. Also, ARF heterodimerization and competition for the binding site may be complementary mechanisms to further tune ARF gene regulation. Overall, these mechanisms can increase the complexity of the auxin response pathway, combining multiple DNA selection mechanisms with the range of members in this transcription factor family, thus determining the final outcome of ARF-mediated gene regulation.



Ancillary Domain

4 The Royal Family of methylation readers 57

- 4.1 Abstract
- 4.2 Introduction
- 4.3 The Royal Family
- 4.4 Discussion

5 In search of an Ancillary Domain function 77

- 5.1 Abstract
- 5.2 Introduction
- 5.3 Results
- 5.4 Discussion

6 Steward Domain: The ARF epigenetic link 95

- 6.1 Abstract
- 6.2 Introduction
- 6.3 Results
- 6.4 Discussion



4. The Royal Family of methylation readers

4.1 Abstract

The ARF Ancillary Domain was originally described after the first crystal structure solution of *Arabidopsis thaliana* ARF1 and ARF5. It shares similarities with Tudor domains, a histone PTM reader module, although the amino acids required for the binding in Tudor domains were not identified in the ARF Ancillary Domain. We have studied the classification of Tudor domains and related structures, known as the “Royal Family”, in order to identify common structural motifs and characteristics of this superfamily. We highlight and discuss the present inconsistencies of the classification and we propose an alternative approach for future classification of newly discovered histone PTM readers with structural similarity to RF.

4.2 Introduction

The cellular context is of high relevance to how auxin response factors select target genes for expression or repression [80]. In contrast to the simple organization of prokaryotic genomic DNA, eukaryotic genomic DNA is highly packed in the nucleus in the form of chromatin [81]. This dense structure allows for the packaging of 125 megabases of DNA in *Arabidopsis thaliana* (about 7-8 centimetres) in the tiny space available in the nucleus [82]. DNA condensation in the form of chromatin prevents the expression of genes, so the binding of transcription factors, including ARFs, to the promoter regions of the target genes require increased accessibility. Chromatin compaction and decompaction is an actively regulated process during gene transcription [83]. The smallest structural element of chromatin is the nucleosome, in which 145-147 base pairs are wrapped around an octamer of core histones [84]. This octamer is formed by two H2A-H2B dimers and a H3-H4 tetramer, for which different variants exist, and provides the scaffold for genetic modulation [46, 85, 86]. The nucleosomes are interconnected by linking regions of the DNA of variable length, which are bound to Histone H1 (linker histone). Contrary to core histones which are only present

in eukaryotes, the linker histone is thought to have originated in bacteria [87]. The DNA structure comprising nucleosomes and linker DNA has been described as “beads on string”.

There are four core histone families, H2A, H2B, H3 and H4, being one of the most conserved histones during evolution [87], although the different types share low sequence similarity. Despite this low sequence similarity, all core histones retain a well-defined core structure, the canonical histone fold [88]. This fold consists of a helix-strand-helix-strand-helix motif complemented with unstructured tail regions at the N-terminus, and in Histone H2A also at C-terminus, representing up to 30% of their length [86]. There are several factors that can result in differences in chromatin composition and dynamics, and thereby lead to spatiotemporal precision in recruitment of specific chromatin components [89]. These factors include histone variant exchange in the nucleosome and covalent posttranslational modifications (PTM) in both the histone tails as well as in the DNA bases.

Dynamic control of the addition, removal and recognition of Histone PTM is precisely regulated by a network of reader proteins and modifying enzymes. The histone code hypothesis, proposed nearly 20 years ago [90, 91], states that isolated and/or combined histone PTMs function as recruiting points for proteins with various functionalities or directly alters the physical chromatin structure [92, 93]. Domains within chromatin readers that confer selective binding to histone marks have been identified among many nuclear proteins. The first domain discovered conferring such selectivity was the chromo box, now called the Chromodomain (CHRomatin Organization MODifier domain) [94]. After this initial discovery, many other chromatin reader domains were identified in a variety of genes, highlighting the importance of histone PTM addition, removal and recognition. The recruitment of these reader proteins results i) in further modification of histones, adding or removing PTMs; ii) in engagement of the transcriptional machinery for gene transcription; and/or iii) in chromatin compaction and decompaction, regulating gene transcription [92, 95]. Thus, the possibility of crosstalk between signalling pathways may depend on competition in adding or removing different PTMs from chromatin by chromatin remodellers involved in different pathways, which would integrate the signals from the cellular status and environmental triggers to provide a coordinated response. Consequently, the plasticity of the chromatin chemistry and structure is a critical regulator of cellular reprogramming, determining which genes are silenced or activated [96].

Activity of Auxin Response factors has been linked to chromatin dynamics. Previous reports suggested that AtARF5 binds through its middle region to Brahma [46, 48], a SWI/SNF chromatin remodeller complex. The C-terminus of Brahma contains a bromodomain that is able to bind histones H3 and H4 [97]. However, no direct connection between ARFs and chromatin dynamics has yet been observed. The first structure of AtARF1-DBD and AtARF5-DBD [39] revealed that DBD is formed by 3 subdomains, named as B3 or DNA-binding subdomain, Dimerization Domain and Ancillary Domain. The B3 and dimerization domains have a very clear function in the structure, but the function of the ancillary domain is not yet understood [27]. The **Ancillary Domain** (AD) [26, 32, 33, 35, 39, 43, 56], or flanking domain [33, 35, 38] is the last structured region before the middle region, comprising the final 80 amino acids of the ARF-DBD. The fold of the AD is similar to that of PHF20-Tudor domain, a member of the so-called **Royal Family** (RF) [39, 98]. Royal family members are known for their ability to bind several posttranslational peptide modifications, especially those located on histone tails. Initial structural analyses of the ARF-ADs discarded the PTM-binding function because the hydrophobic cage (HC) that recognizes PTM was occluded in the structure [32, 33, 39, 43, 56]. Since then, the ARF AD did not attract attention and a function has not yet been attributed to it [32].

In this chapter, we review the present structural knowledge on the Royal Family protein domains with the aim of identifying a possible evolutionary relationship between the ARF Ancillary Domains and the Royal Family. We give an overview of the current classification of the main families belonging to the “Royal Family” superfamily and the structural particularities of each member. We establish the minimal structural elements that define the RF superfamily and classify protein families according to the presence of these elements in their structures. Then, subfamilies could be grouped based on their sequence similarity to identify differences between members of the same family. For this purpose, we will use the families related to the RF found in Pfam, Interpro and SCOP databases. In addition, we review other non-Royal Family histone methyllysine readers in order to have the complete landscape of structures with methyllysine and methylarginine binding properties. We retrieved structures of each class to define structural characteristics, which has as far as we know not been considered in detail for the present classification. We developed Python scripts to retrieve the structures from PDB, superpose them, get the HC sequence conservation and finally classify each of them on the basis of their structure. With this approach, we compared the classification of each family in several structure classification databases in order to detect possible inconsistencies. All the information gathered in this search can be found in the supplementary material (supplementary tables 9.8, 9.9 and 9.10) and the classification criteria for each family is summarized in the discussion, which we hope will be useful for the classification of new methyllysine and methylarginine binding protein structures.

4.3 The Royal Family

The Royal Family is a protein superfamily composed of domains of methyllysine and methylarginine readers, structurally and functionally related with the Tudor domain [99, 100]. Tudor domains were originally discovered in the protein Tud encoded by the *tudor* gene in *Drosophila melanogaster*, where mutations in this gene result in offspring lethality or infertility, hence the reference to Tudor king Henry VIII [100]. Later, as other folds similar in structure and function to Tudor were found, the name of “Royal Family” was coined to consider all these evolutionary related domains [101].

The Royal Family domains usually endow proteins with interaction specificity to modified chromatin [99]. The minimal structure of a Royal Family member consists on a three-stranded antiparallel β -sheet. To this basic structure, other structural motifs are included that tune the binding specificity and affinity, characteristic of each family. All Royal Family domains share a hydrophobic cluster (HC) that is the defining feature responsible for of all methyllysine or methylarginine binding proteins [100]. The HC motif consists of two to four aromatic residues that provide hydrophobic and cation- π interactions with the methylated substrate. The specificity for the recognition of mono-, di- and tri-methylated lysine, as well as symmetric and asymmetric arginine dimethylation relies in the overall hydrophobicity and charge of the pocket residues as well as the size and shape of the binding site [95]. Residues outside the HC provide additional anchor points for recognition and regulation, thereby determining the sequence specificity for other residues of the substrate. As a consequence, highly hydrophobic HCs will tend to interact with trimethylated lysines, whereas the presence of acidic residues in the HC can establish hydrogen bonds with charged lysine or arginine and therefore tend to interact with charged, mono- or di- methylated lysine or arginine residues. The families that belong to the Royal Family of methyl readers are the Tudor domains, PWWP domains, MBT domains and Chromo-like domains [94, 99, 100, 102, 103]. The differences that characterize a protein as belonging to a certain RF subfamily are small, supporting a common evolutionary ancestor. RF members consist of ≈ 60 amino acid, with high structural similarity with SH3 protein-protein interaction domains [99, 104], where each RF member either includes or lacks

elements of secondary structure or has varying positions of homologous elements.

The classification of RF proteins has been primarily done by sequence similarity given the fact that no structural information was available at the time of discovery. As a result, proteins have been classified in distinct families that were later found to share high structural homology. The RF superfamily was defined when the sequence similarity between Tudor and MBT and between MBT and chromo domain became evident. Thus, Tudor and chromo domain were initially linked to MBT [101]. Structural analyses then led to the addition of PWWP to RF for its similarity, making the classification of the RF subfamilies a mix between sequence and structure similarity.

4.3.1 Chromo-like domains

The chromo-like family of domains is composed of three families related by sequence and structural homology: The Chromodomain [105], the Chromo Barrel Domain [106] and the Chromo Shadow Domain [94]. Some authors [104] consider Chromo-like as a superfamily, although they are annotated to belong to the “Royal Family” superfamily in the SCOP database. For the purpose of establishing a consistent classification hierarchy, we will treat Chromo-like as a family of the “Royal Family” superfamily, which are further divided into subfamilies.

Chromodomain

The chromo domain family can be found in databases as Chromo (pfam, PF00385), Chromo domain family (SCOP, 4000375) or as Chromo domain (Interpro, IPR023780). The canonical chromodomain consists of around 60 amino acids arranged as three antiparallel β -strands and a C-terminal α -helix packed against it, adopting a half barrel structure. This topology represents the simplest form of a Royal Family member [99] and is similar to that found for SH3 domains [104, 107], where the chromodomain lacks the first strand of the SH3-like beta-barrel. Upon binding, the recognition peptide forms the lacking first β strand, resulting in a SH3-like barrel [107, 108] (Figure 4.1) for example in the structure of 1KNA. 3_{10} helices may be present in the loops connecting $\beta 2$ - $\beta 3$ and $\beta 3$ - $\beta 4$. The typical chromodomain C-terminal α -helix is located after $\beta 3$, packing against the β -sheet. While this description applies to most chromodomains, other more complex folds are also classified as chromodomains in databases. The number of β strands in these proteins increase from the typical three to five, where the extra two β strands occur N-terminal to $\beta 1$ and C-terminal to $\beta 3$, respectively. As we will describe below, this fold consisting in 5 antiparallel β -strands is more consistent with a Chromo Barrel domain (also known as Tudor-knot).

The hydrophobic cage responsible for the methyllysine recognition is formed by at least three aromatic residues, located in strands $\beta 1$ (Nter), $\beta 2$ (Cter), and in the loop connecting $\beta 2$ to $\beta 3$ [99]. In most cases, HC residues in $\beta 2$ and $\beta 3$ -4 loop are tyrosines, and the $\beta 3$ residue is a tryptophan. This tryptophan is substituted by a tyrosine in 2RNZ, a structure classified as chromodomain in the literature but as the related Tudor-Knot (Chromo Barrel) by pfam and interpro. Other residues close to the HC may contribute to the stabilization of the bound methylated residue. A polar residue, mainly glutamate or glutamine, is located in the 3_{10} helix after $\beta 3$ and may assist in stabilization of the methylated residue.

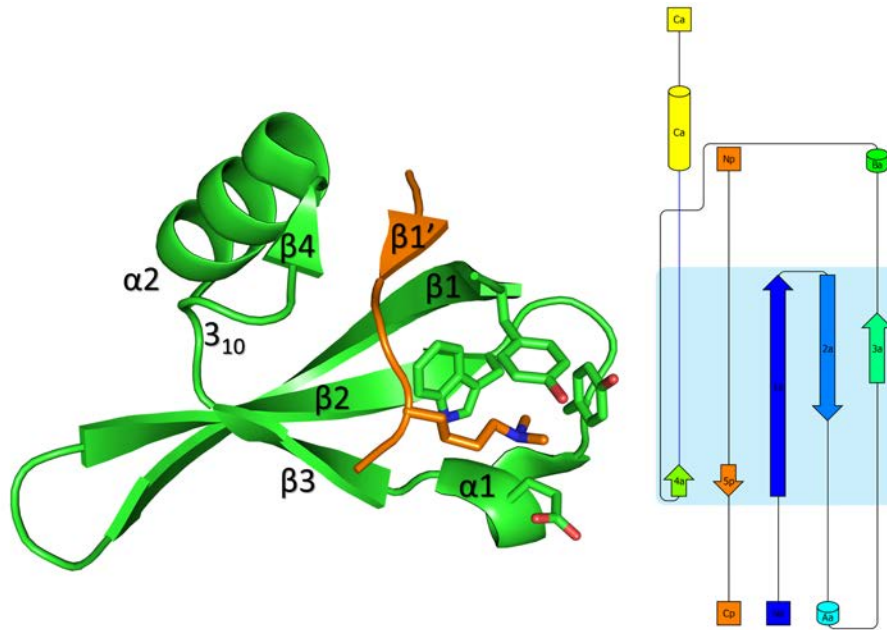


Figure 4.1: Structure of a canonical chromodomain. Left panel, cartoon representation of the *Drosophila melanogaster* HP1 chromodomain structure (green) in complex with histone H3 tail peptide containing dimethyllysine 9 (orange). This structure displays the typical chromodomain arrangement of 3 antiparallel β strands followed by a C-terminal α -helix (PDBid: 1KNA [108]). The interacting dimethylated lysine and the residues composing the Hydrophobic Cage are displayed as sticks. Right panel, 2D layout of the structure, with the interacting histone peptide coloured in orange.

Chromo Barrel domain

The general structure consists of five antiparallel β strands, a $\beta 4$ - $\beta 5$ 3_{10} helix and an optional but frequent $\beta 3$ - $\beta 4$ 3_{10} helix. N and C terminal α helices may be present in this domain and the loop connecting $\beta 3$ - $\beta 4$ may contain various structural elements. Strands $\beta 2$ - $\beta 4$ are equivalent to the chromodomain $\beta 1$ - $\beta 3$ core strands, and the first chromo barrel domain β -strand shares position with the β -strand formed by the lysine-methylated peptide target in the chromo domain. This arrangement prevents the Chromo Barrel Domain to recognize the histone peptide in a similar way as the chromodomain, as the free space in chromodomain for peptide binding is occupied by $\beta 1$ in Chromo Barrels. The characteristic C-terminal α -helix of the chromo domain is substituted by $\beta 5$ in the Chromo Barrel domain [99, 101, 106, 109] and at the N- and C-terminus a variety of structural elements as loops, β -strands and α -helices may be found. In this type of domain, the interaction with the substrate occurs at the open end of the barrel that contains the hydrophobic cage (Figure 4.2)

Despite the similar name, the structure of the Chromo Barrel domains differ greatly from chromo domains [106] and structurally resemble Tudor or MBT domains more [99]. While some structures belonging to this family are classified as chromo barrel domains in SCOP (Chromo Barrel domain (4003139)), they are classified as Tudor-knot by pfam and interpro databases (PF11717 and IPR025995, respectively), highlighting the similarity with Tudor fold, despite sharing sequence homology to chromodomains. Tudor-knot domains were defined by their RNA-binding ability and by the presence of N- and C-terminal β strands ($\beta 1$ and $\beta 5$) that interact with each other, establishing a “knot” of the structure [111]. This suggests that Chromo Barrel domains and Tudor-

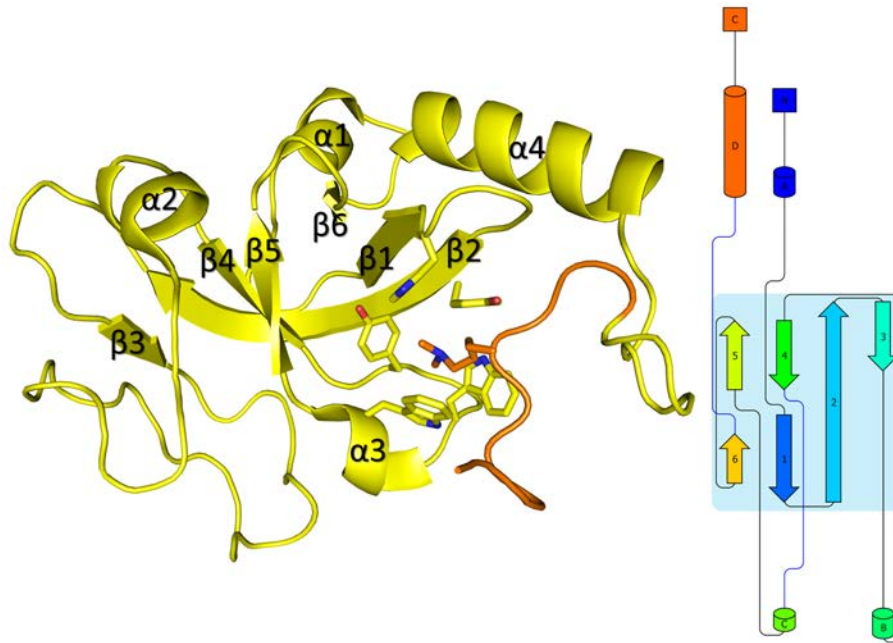


Figure 4.2: Structure of a Chromo Barrel Domain. Left panel, cartoon representation of the solution structure of the EAF3 Chromo Barrel Domain (yellow) bound to histone H3 tail peptide containing dimethyllysine 36 analogue (orange) displays the typical arrangement of 5 core antiparallel β strands followed by a C-terminal α -helix (PDBid: 2K3Y [110]). The interacting dimethylated lysine and the residues forming the Hydrophobic Cage are shown as sticks. Right panel, 2D layout of the structure.

knot domains structures cannot be distinguished structurally. For consistency, we propose to merge both definitions under the Chromo Barrel domain.

The hydrophobic cage observed in the analysed Chromo Barrel structures is formed by at least four residues of the aromatic or hydrophobic type, to which a polar amino acid (N or E, for example) is added in some cases, which helps to stabilize ligand binding. The hydrophobic and aromatic residues are found at the C-terminal end of $\beta 1$ and $\beta 3$, at the N terminal end of $\beta 2$, in the loop connecting $\beta 3$ -4 and in the 3_{10} helix found in $\beta 3$ -4 loop. Thus, the HC of the Chromo Barrel structure consists of more residues than the HC of the chromodomain, and one of the residues is provided by $\beta 1$, which is not present in the chromodomain. The polar amino acid, when present, may substitute the residues at the beginning of $\beta 2$ or in the 3_{10} helix.

Chromo Shadow domain

The general structure of the Chromo Shadow domain adopts a similar fold as the canonical chromodomain, as it consists on a 3 stranded antiparallel β -sheet onto which an α -helix is bound (Figure 4.3). Interestingly, the SCOP database does not contain all the structures analysed, and those listed are annotated to be Chromo domain family (4000375). In contrast, all proteins are listed as Chromo Shadow in pfam and interpro (PF01393 and IPR008251, respectively). The Chromo Shadow domains lack a proper HC. Thus, several of the analysed Chromo Shadow domains contain hydrophobic residues in some but not all the corresponding chromodomain HC positions. In addition, the N-terminus of Chromo Shadow domains usually folds as an α -helix that is oriented antiparallel to the C-terminal helix, which blocks the hydrophobic cage. The

absence of the HC and its inaccessibility caused by the first α -helix suggest that these domains would not be able to bind methyllysine as in chromodomains. Chromo Shadow domains always appear in conjunction with a chromodomain [94, 112]. Structural and biochemical studies on Chromo Shadow domains suggested that this domain is able to homodimerize, and thereby allow Chromodomains to oligomerize and condense chromatin [113].

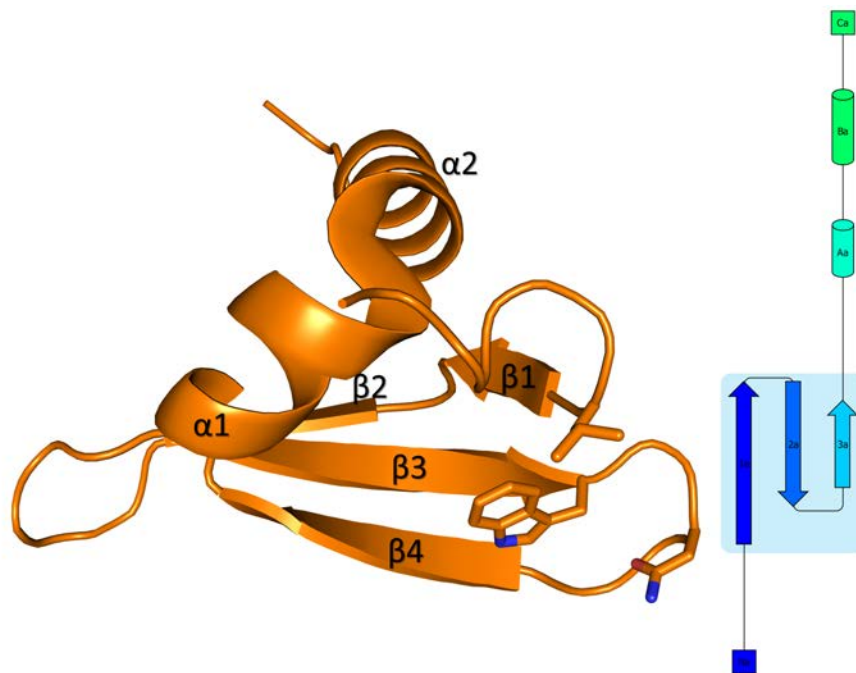


Figure 4.3: Structure of a Chromo Shadow Domain. Left panel, cartoon representation of the structure of a monomer of the Chromo Shadow Domain from *Schizosaccharomyces pombe* swi6 protein (orange) (PDBid: 1E0B [113]). The residues that share sequential and geometrical position with the residues forming the Hydrophobic Cage in chromodomains are shown as sticks. Note that these residues are found at the surface of the Chromo Shadow domain, and do not form a cage. Right panel, 2D layout of the structure

Some of the observations made in our analysis questions the validity of the classification of the Chromo Shadow domain as a separated subfamily within chromo-like domains based on similar folding and sequence conservation [113]. First, Chromo Shadow domains always appear in combination with a chromodomain, suggesting that this combination should be considered as a functional unit. Secondly, the absence of an HC suggests that the Chromo Shadow domains are functionally different from proteins of the RF superfamily since binding of methylated peptides to a functional HC is a characteristic of all RF domains. Taking all of this in consideration, we believe that the combination of the chromodomain and Chromo Shadow Domain should be classified as a single domain within the chromo-like family, separate from the chromodomain subfamily. In fact, this configuration is similar to the extended Tudor domains, reviewed later, where function is dependent on a combination of a functional Tudor and an SN-like domain, which, similar to Chromo Shadow Domains, has an RF fold but lacks an HC and the ability to bind methylated substrates (Figure 4.3).

4.3.2 Tudor-like domains

Tudor-like domains are an important protein-protein interaction module present in proteins with diverse functions related with chromatin signalling, such as gene transcription, epigenetic inheritance, RNA metabolism and DNA damage response [100]. The ≈ 60 amino acids of the core of the Tudor-like domain fold into a strongly bent antiparallel four/five-stranded β -barrel [100, 114]. Three of these antiparallel β -strands share position with other members of Royal Family, suggesting that all originate from a common ancestor [99]. In fact, the residues forming the HC in Tudor-like domains and other members of the Royal Family are generally conserved. The proteins containing a Tudor-like domain are the only histone readers that are able to recognize methyl-arginine substrates, all other known histone readers, both from RF as well as other superfamilies, are exclusive methyllysine readers [99, 100].

Tudor-like domains are currently classified into four different types according to their function. Tudor-like domains can be present as single Tudor domain, or as multiple Tudor domains. The latter are subdivided into Tandem and Hybrid Tudor domains and are present in proteins as two consecutive Tudor folds occurring in the same polypeptide chain. From these two domains, only one is active. The last subfamily of Tudor-like domains are the extended Tudor domains, which contain a functional Tudor domain combined with a non-functional SN-like domain. While Tandem and Hybrid Tudor domains are exclusively methyllysine binders, extended Tudor domains are known methylarginine readers, whereas Single Tudor domains with methyllysine or methylarginine binding properties are known [99, 100]. With all this variability and the elaborated β -sheet core compared to other Royal Family members, we consider the Tudor-like domains as the most complex family inside Royal Family.

Single Tudor domain

The single Tudor domains discovered up to date are able to recognize both methyllysine- and methylarginine-containing peptides [100, 114]. All the analysed structures, without exception, contain an antiparallel β strand barrel-like structure, consisting of four or five strands and a 3_{10} helix between $\beta 4$ - $\beta 5$. $\beta 5$ may be very short or absent, but in all cases the region comprising the position of the $\beta 4$ - 3_{10} helix and $\beta 5$ is covering and closing the barrel on one side (Figure 4.4). The single Tudor domain represents the basic layout to which additional structural elements in other Tudor-like domains are added to create more complex forms of Tudor-like domains and thereby tune the activity. Although the structure is highly conserved in the analysed single Tudor domains, they are integrally annotated as Tudor domains in the Interpro database only (IPR002999). In SCOP and pfam databases, Tudor and SMN, a family of proteins containing single Tudor domains, are mixed with non-Tudor domains in the classification.

The HC of Single Tudor domains is formed by 4 or 5 hydrophobic residues, many being aromatic residues (at least 3). The cage is complemented with one or two polar amino acids to stabilize the charge of the methylated lysine or arginine bound to the cage. The cage residues are located at the C-terminus of $\beta 1$ and $\beta 3$ and at the N-terminus of $\beta 2$ as in the Chromo Barrel domains. Tudor domain HCs are more hydrophobic compared to those found in chromo-like domains, as hydrophobic residues at the N-terminus of $\beta 4$ or located in the $\beta 3$ - $\beta 4$ loop may also participate in the HC formation. These residues alternate between a hydrophobic and polar nature which may further stabilize the interacting methylated arginine/lysine.

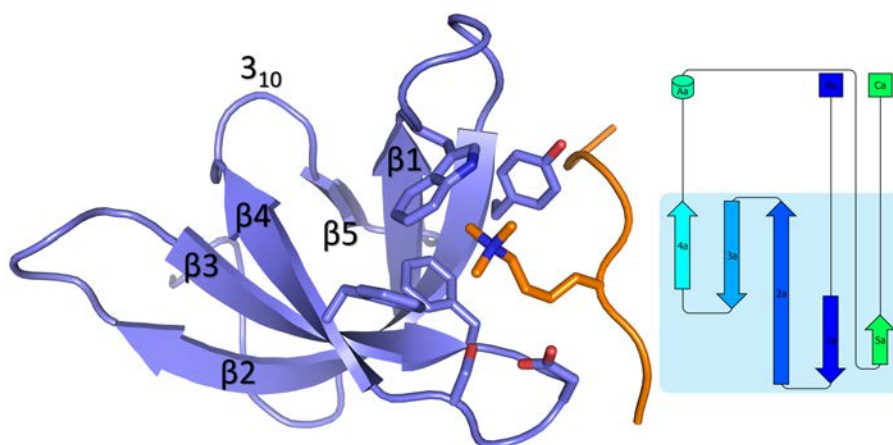


Figure 4.4: Cartoon representation of the structure of the single Tudor domain of human PHF1 protein. Left panel, structure of the single Tudor domain (blue) (PDBid: 4HCZ [115]) displays a five-stranded β barrel, typical of Tudor domains. Four aromatic residues, shown as sticks, are located in strands $\beta 1$ - $\beta 4$ and form an aromatic cage that lock in the trimethylated lysine 36 (orange sticks) from the H3 peptide (orange). Other acidic residues of the domain, also shown as sticks, aid the hydrophobic cage to bind the H3 tail peptide. Right panel, 2D layout of the structure.

Tandem Tudor domain

The Tandem Tudor motif consist of two or more consecutive repeats of Tudor-like folds, connected by random coiled loops (Figure 4.5) The structures of the individual domains are very similar to that of the single Tudor domains. Most structures also contain helical turns in the loop $\beta 2$ - $\beta 3$ or in the loops connecting the domains. These helical turns are placed at the opposite site of the HC with respect to the barrel-like structure, covering the HC and presumably stabilizing the whole structure. In all the structures analysed, both Tudor domains contain a complete HC at similar positions as in the single Tudors that would allow both domains to function as methyllysine reader, conferring to the protein the ability to recognize multiple amino acid modifications in a peptide. However, some of the domains in the repeat are in closed conformation incompatible with recognition of the methylated peptides. Furthermore, the residues located on the loops connecting both Tudor domains or in the $\beta 4$ of the domain may block the HC cage in the closed conformation domain. It has been proposed that occlusion by nearby residues blocks the function of the hydrophobic cage in the closed conformation, but the biological function and evolutionary significance of these obstructed Tudor domains remains unknown [116]. Flexibility of the connecting loops or $\beta 4$ may allow for structural reorganization that activates the HC in the inhibited Tudor domains.

The plant Agenet Domains are a group of RF domains that share the structural and organizational characteristics of Tandem Tudor domains. Sequence searches originally showed that Agenet domains are a distant relative of the Tudor domains. In consequence, this group of Tudor homologs found in plants were named plant Agenet from the Plantagenet dynasty of English monarchs [101]. The Agenet domain was initially found as a plant-specific "Royal family" module, but its function remained unknown [118]. Recent bibliography shows that Agenet domains mostly appear in tandem and have the ability to bind methylated lysines from histone peptides [109, 118]. A careful look of the structure show that Agenet domains belong to the Tandem Tudor domain class. In fact, there are studies that independently consider these domains as Agenet or Tandem Tudor domains. For example, the N-terminal domain of the Human Fragile X mental retardation protein is classified as either Tandem Tudor [119] or Agenet [120] in different studies. This is probably a consequence of

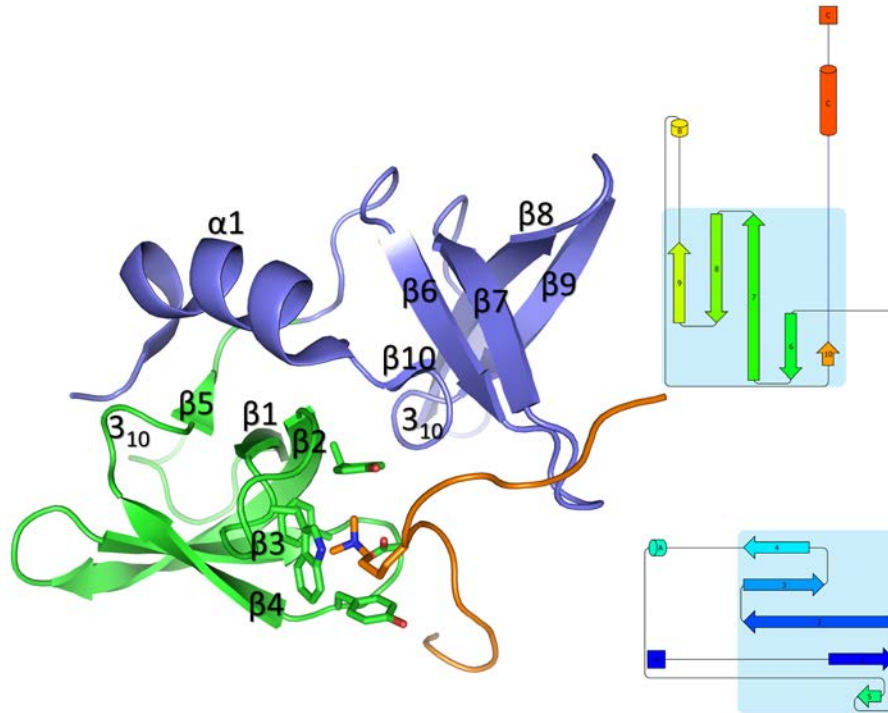


Figure 4.5: Cartoon representation of the structure of the tandem Tudor domain of human 53BP1. Left panel, structure of the tandem Tudor domain (PDBid: 2LVM [117]) shows two domains in tandem (green and blue) contained in the same polypeptide. The N-terminal domain, which binds the H4 dimethylated lysine 36 peptide (orange), is shown in green and the C-terminal domain is shown in blue. Both domains display the typical five-stranded β barrel Tudor fold. Four aromatic residues in the N-terminal Tudor domain, shown as sticks, are located in $\beta 2$, $\beta 3$ and $\beta 4$ strands and the loop connecting $\beta 1$ - $\beta 2$ and form an aromatic cage that lock the dimethylated lysine 36 from H4 peptide (orange sticks) into place. An aspartic acid residue located in the loop $\beta 3$ - $\beta 4$ of the domain, also shown as sticks, aids the hydrophobic cage to bind tightly to the H4 tail peptide. Right panel, 2D layout of the structure.

the independent discovery of the Agenet and the Tandem Tudor domain groups before resolution of Agenet structures, which revealed the structural and sequence similarity between them. With the aim to maintain the “Royal Family” spirit of this protein superfamily, we propose that the plant Agenet and the Tandem Tudor domain can be classified as the Tandem Agenet domain.

Hybrid Tudor domain

The structure of hybrid Tudor domains generally consists of two interdigitated Tudor domains where two of the four antiparallel β -strands ($\beta 3$ and $\beta 6$) that make up the core β sheet are long β strands that are shared by the two domains (Figure 4.6). The first domain is formed by the N-terminal $\beta 1$, $\beta 2$ and part of $\beta 3$, and by the C-terminal $\beta 6$, $\beta 7$ and $\alpha 3$. For this reason, we refer to it as the **N/C Tudor domain**. The first two antiparallel β strands $\beta 1$ and $\beta 2$ in the N/C Tudor domain are in the same configuration as in canonical Tudor domains. $\beta 3$ is a long strand, which is shared with the other Tudor domain. The second Tudor domain is formed by the central $\beta 3$ -6 and $\alpha 1$ - $\alpha 2$. As the structural elements that form this second Tudor domain are found in the central part of the amino acid sequence, we will refer to it as the **central Tudor domain**. Thus, the central Tudor domain can be considered an insertion into the N/C Tudor domain. The C-terminal part of

$\beta 3$ occupies a position in the central Tudor domain that is equivalent to $\beta 3$ in canonical Tudor domains. Then, $\beta 4$ of the Hybrid Tudor domain is placed in central Tudor domain in the equivalent position to a $\beta 4$ in canonical Tudor domains, followed by an α -helix replacing the typical 3_{10} helix of the $\beta 4$ - $\beta 5$ turn in other Tudor domains. After this helix, another α -helix replaces the $\beta 5$ strand found in Tudor-like domains. The position of these two α -helices is equivalent to those observed in Chromo Shadow Domains, probably due to the evolutionary relationship of Chromo- and Tudor-like domains, but in the hybrid Tudor they are considerably shorter. These two α -helices are closing the barrel-like structure at the opposite end of the HC opening at the central Tudor domain. This configuration is not present in N/C Tudor domain, leaving a barrel structures with the two sides open. $\alpha 2$ helix in the central Tudor domain is followed by strand $\beta 5$. $\beta 6$ is the next structural element present, which is again a long β strand, antiparallel to strand $\beta 3$, that connects the central Tudor to the N/C Tudor. The last $\beta 7$ and $\alpha 3$ finalize the structure of the N/C Tudor domain.

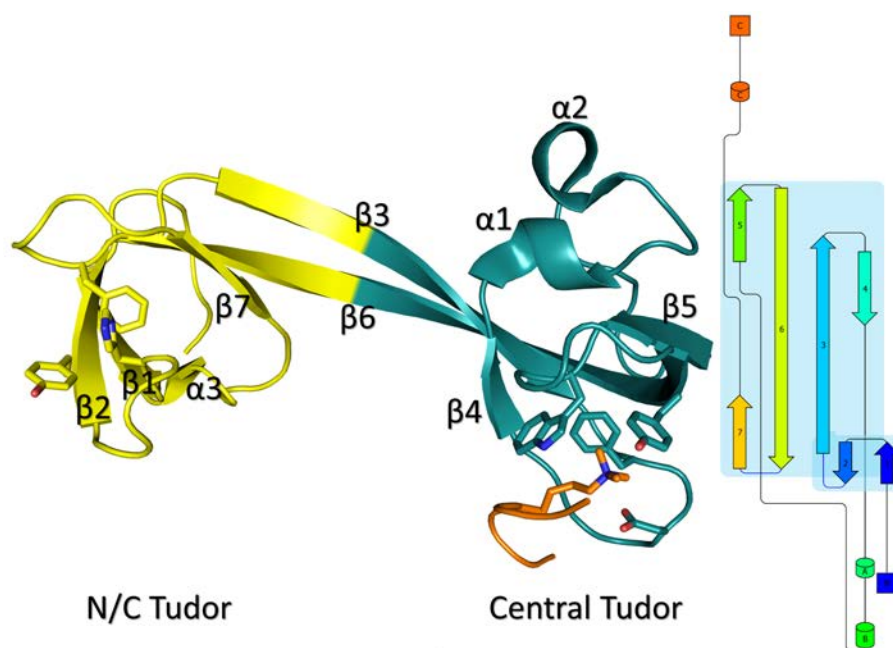


Figure 4.6: Cartoon representation of the structure of the hybrid Tudor domain of human JMJD2A histone demethylase in complex with the Histone H4 peptide trimethylated at lysine 20. Left panel, the structure of the hybrid Tudor domain (PDBid: 2QQS [121]) shows two interdigitated domains. The N/C Tudor domain is shown in yellow and the central Tudor domain, which binds the methylated histone peptide, is shown in cyan. Both domains display the typical β -barrel Tudor fold, with the particularity that the central strands corresponding to β -3-b6 are shared between the two domains. The central Tudor binds the trimethylated lysine 20 from H4 peptide (orange sticks). Right panel, 2D layout of the structure.

In the structure, the interacting peptide is bound to the central Tudor domain. The HCs of the two Tudor domains differ considerably from the canonical HC found in other Tudor domains. The analysed structures of the N/C Tudor domain have only two to three of the four to five hydrophobic residues that are generally present in the HC of canonical Tudor domains, and the HC of the central Tudor domain contain 3 hydrophobic residues in all the analysed structures. This difference with canonical Tudor domains suggests a different specificity and affinity. There is no evident reason for that the N/C Tudor domain does not bind methylated peptides.

Extended Tudor domain

This motif consists of two domains, one corresponds to a Tudor domain, the other to an SN-like domain, which also is similar in fold to a Tudor-like domain (Figure 4.7). It is the largest motif from the Tudor-like family and spans ≈ 180 amino acids. The overall fold can be considered as a canonical Tudor domain, preceded by a long α helix, inserted into a *Staphylococcal nuclease-like domain* (SN-like domain). The point of insertion is between the second β strand and third β strand of the SN-like domain. The Tudor domain consists of the five antiparallel β strands forming the twisted β barrel and the 3_{10} helix in the $\beta 4$ - $\beta 5$ loop. The HC is formed by 4 aromatic residues and a polar asparagine residue, all at the canonical positions. A small α -helix is present following $\beta 5$, which is antiparallel to the $\alpha 1$ helix and packed against the Tudor domain. The SN-like domain also displays a Tudor-like fold, although the hydrophobic cage is not present [100, 122, 123]. Existing reports suggest all three structural elements of the motif (the canonical Tudor domain the long $\alpha 1$ helix and the SN-like domain) are required for function [100, 122, 123].



Figure 4.7: Cartoon representation of the structure of the extended Tudor domain of human SND1 protein. Right panel, the canonical Tudor domain is shown in purple, and the *Staphylococcal nuclease-like domain* (SN-like domain) is shown in pink (PDBid: 3OMC [122]). The peptide binds to the hydrophobic cage of the Tudor domain. This cage, shown as sticks, is composed of 4 aromatic residues complemented with an aspartate and asparagine. The symmetrically dimethylated arginine-containing peptide, coloured in orange, binds to this cage. The numbering of the secondary structure elements is based on the order of appearance in the Tudor-like domain. Right panel, 2D layout of the structure.

MBT domain

The MBT domains take their name from the original discovery of three repeats of the MBT fold in a *Drosophila melanogaster* gene called *lethal (3) malignant brain tumour*. The general structure of this domain (Figure 4.8) comprises two or more repeats of the basic MBT fold, a fold consisting of a core formed by a five stranded antiparallel β -barrel ($\beta 1$ - $\beta 5$) with the exact same strand topology and similar relative positions as found on Tudor-like domains. Consecutive domains interact, where the N-terminal part of one domain interacts with the β sheet core of the next domain [99, 101]. The core harbours an hydrophobic/aromatic cage, which allows this domain to recognize mono- or dimethylated lysins with high affinity and specificity ($K_D \approx 5 \mu\text{M}$) [101]. The HC of the domains contain at least 3 aromatic rings, complemented in some cases with histidine, prolines and/or polar residues such as an aspartic acid. The HC residues are located at the end of $\beta 3$, in $\beta 4$ and in the loop connecting $\beta 3$ - $\beta 4$, and therefore are positioned similarly as in Tudor-like domains. The auxiliary residues are located in $\beta 1$ and $\beta 2$ when present.

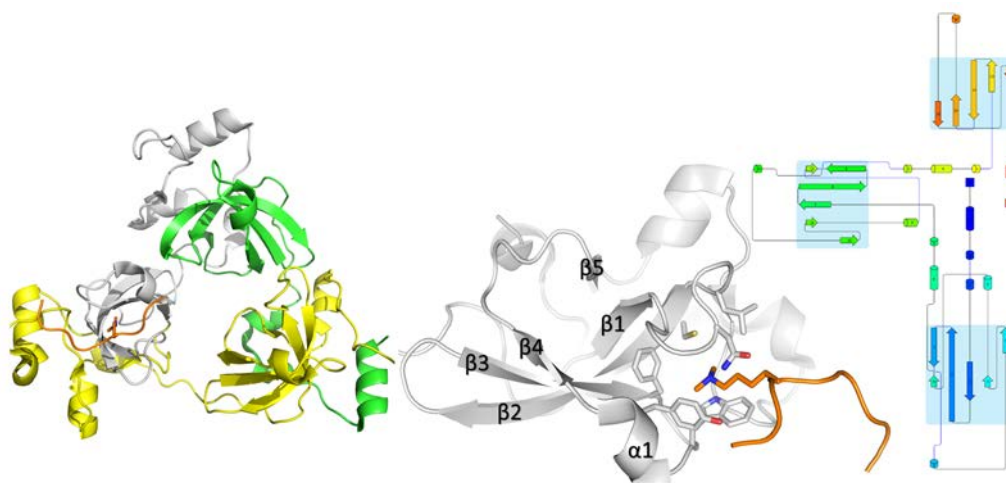


Figure 4.8: Cartoon representation of the structure of the Human L3MBTL1 in complex with H4K20Me2 peptide. Left panel, the complete structure of Human L3MBTL1 is shown as cartoon (PDBid: 2PQW [124]). The first MBT repeat is shown in green, the second in grey and the third in yellow. Note that N-terminal helices of each MBT repeat interact with the previous MBT, stabilizing the whole structure. The H4K20Me2 peptide interacting with repeat two is coloured in orange. Central panel, a single repeat of the MBT domain is shown in grey, where the residues of the hydrophobic cage are shown as sticks. The interacting H4K20Me2 peptide is shown in orange, where the dimethylated lysine 20 is shown as sticks. Right panel, 2D representation of the layout of Human L3MBTL1.

It is possible that not all the MBT repeats found in an amino acid sequence function as methyllysine binders, as the HC of some of the repeats contain little or no hydrophobic residues. As in other Royal Family members, strands $\beta 3$ and $\beta 4$ are often connected by a 3_{10} helix, but the 3_{10} helix between $\beta 4$ - $\beta 5$ found in Tudor domains is absent in MBT domains [99, 101]. The domains are preceded by variable N-terminal loops, 30 to 50 amino acids in length, containing 3_{10} - or α -helices that pack against the opposite site of the β barrel with respect to the HC of the neighbour MBT. Part of the N-terminal loop also folds as a short β -strand, which interacts antiparallel to $\beta 2$ of the core of the preceding MBT. This structurally stabilizes the relative position between domains.

PWWP

The PWWP domain was originally discovered through a Pro-Trp-Trp-Pro motif in the *Wolf-Hirschhorn syndrome candidate 1 protein* (WHSC1) [125]. It was later found to be a conserved domain that spans 100-130 residues. All the PWWP domains analysed contain three distinct structural regions: a canonical Royal Family β -barrel, a C-terminal α -helical bundle and finally, an insertion between the second and third strands of the β -barrel (Figure 4.9), which is variable in length and in secondary structure. The C-terminal α -helical bundle is formed by at least two and up to four C-terminal α -helices that pack against $\beta 1$ - $\beta 2$ of the canonical Royal Family barrel.

The typical Royal Family barrel is formed by 5 antiparallel β strands, where a 3_{10} helix is located between $\beta 4$ - $\beta 5$, similar to single Tudor domains. The PWWP domain recognizes methyllysine as is the case for most Royal Family members. The binding affinity to the substrates is weak, with K_D s in the millimolar range [99, 127]. The residues forming the HC are located in the $\beta 1$ - $\beta 2$ loop, at the beginning of $\beta 2$, at the end of $\beta 3$ and in the $\beta 3$ - $\beta 4$ loop. The HC consists

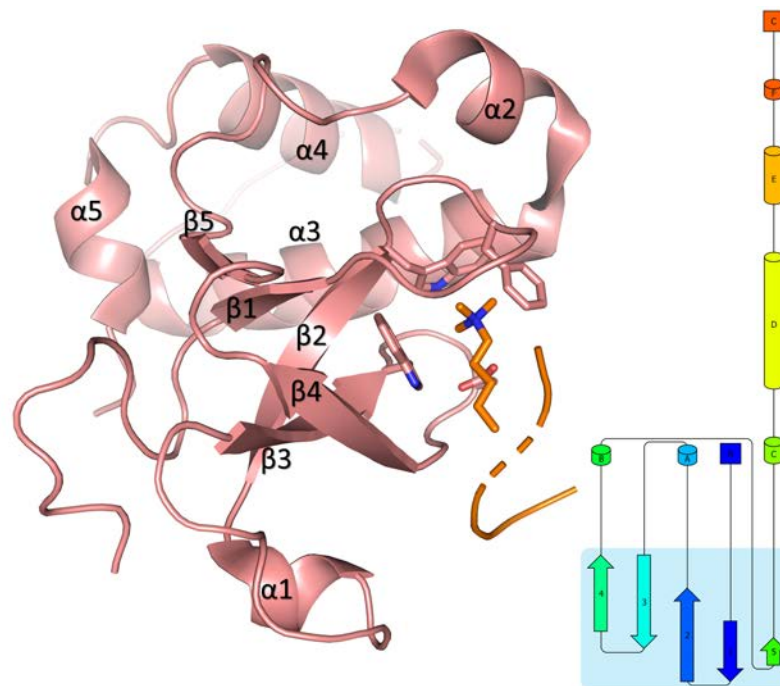


Figure 4.9: Cartoon representation of the structure of the PWWP domain of the Human DNA methyltransferase 3B in complex with H3K36Me3. Left panel, the structure of the PWWP domain, shown in pink (PDBid: 5CIU [126]) contains the typical Tudor fold complemented with several α -helices. The residues composing the hydrophobic cage responsible of H3K36Me3 binding are shown as sticks. The Histone H3 peptide is shown in orange, and the trimethylated lysine 36 is shown as sticks. Right panel, 2D layout of the PWWP domains structure of the Human DNA methyltransferase 3B.

of two or three aromatic and hydrophobic residues, which are complemented by proximal acidic residues. The characteristic PWWP motif that names this Royal Family member is located in the $\beta 1$ - $\beta 2$ loop and functions as a nexus between HC and the CTerminal helices. The exact motif may differ, as PWWP, SWWP and PHWP variants have been found in the analysed structures. However, the motif was located in the above-mentioned position in all cases.

4.3.3 Non-RF Histone methyllysine readers

There are proteins that are methyllysine readers and also use a hydrophobic cage for recognition of the methyllysines but are not considered members of the Royal Family, as they are structurally and sequentially non-homologous. We will review the main characteristics of these non-RF domains in as they exhibit similarities with Royal Family and ARF ADs.

PHD fingers

The PHD (plant homeodomain) finger is a ≈ 50 -residue motif with the ability to recognize lysine residues, mainly found in proteins involved in eukaryotic transcription regulation. The domain folds into an interleaved zinc finger, consisting of a core with a two-stranded antiparallel β -sheet [99, 102, 128, 129, 130]. A characteristic feature in its sequence is a conserved $Cys_4 - His - Cys_3$ (or, less commonly, $Cys_4 - His - Cys_2 - His$) which is responsible for zinc ion binding. PHD fingers are classified in two types based on the presence of a hydrophobic cage or a cage of acidic residues. While the overall structure remains very similar, this difference represents the ability of the PHD finger to bind trimethylated or non-methylated lysine [128].

The PHD fingers of the two subclasses exhibit similar binding affinity for H3K4me3 and me0, respectively, in the high nM to low μ M range [99, 129]. The hydrophobic cage of PHD's that recognizes trimethylated lysines consists of one to four aromatic residues located on the two-stranded antiparallel β -sheet, where additional structural elements in the vicinity may contribute to the HC [99, 102, 131]. A tryptophan residue is always present in the hydrophobic cage of all known methylated histone-recognizing PHD fingers [128]. In all the cases analysed, the HC is not as hydrophobic as in RF members, differentiating the HC of PHD domains from RF domains. The histone peptides bind in a similar mode to that found on chromodomains, where the histone peptide forms a β -strand that packs against the first β -sheet of the PHD, thus adding a strand to the two-stranded β sheet, establishing a three-stranded antiparallel β -sheet [131] (Figure 4.10).

WD40

The WD40 fold consists of 7-8 repetitions of a 4 stranded antiparallel β -sheet subdomain that self-associate to form a β -propeller [99, 133, 134, 135], where the consecutive antiparallel β sheets arrange into a circular structure (Figure 4.11). The first N-terminal β strand found in the WD40 sequence is interacting with the last three C-terminal β strands of the last repetition, which functions as a locking mechanism that preserves the circular shape. Although it is theoretically possible that a β -propeller can be formed by 4 to 8 WD40 repeats [136], structural confirmation has been found only for β -barrels formed by 7 and 8 WD40 repeats [134].

The defining characteristic that classifies β -propeller proteins as part of the WD40 family is the presence of Gly-His (GH) dipeptide at 11-24 residues from its N-terminus and Trp-Asp (WD) dipeptide at the C-terminus of each repeat, although each repeat of the β -propeller can vary between

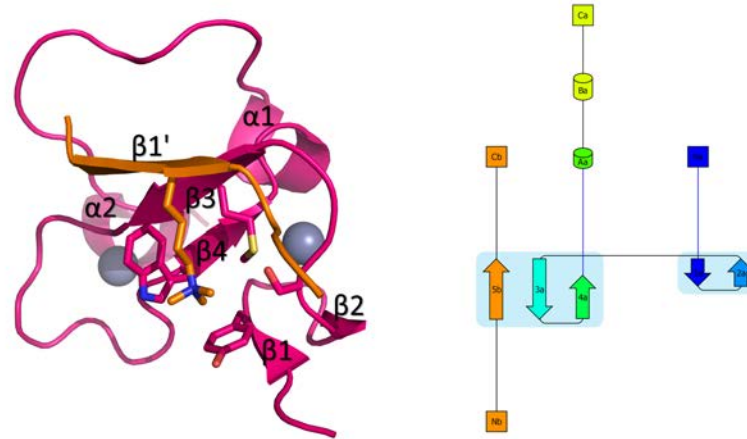


Figure 4.10: Cartoon representation of the structure of the PHD finger of the tumour suppressor ING2 of *Mus musculus* in complex with H3K4me3. Left panel, the plant homeodomain of ING2 is shown in pink, while the interacting trimethylated lysine 4 Histone H3 peptide is shown in orange (PDBid: 2G6Q [132]). Zn^{2+} ions are shown as grey spheres. The residues of the hydrophobic cage that participate in the binding and the trimethylated lysine residue are shown as sticks. Right panel, 2D layout of the structure of the PHD finger of ING2. The interacting peptide is also coloured in orange.

40 to 60 residues in length, and can have variable sequences. The name of the WD40 family derives from the presence of the characteristic WD dipeptide at the C-terminus and a typical length of ≈ 40 residues on each repeat [99, 133, 134, 135, 137].

The primary function of WD40 domains is related to protein-protein interaction, where the β -propeller structure gives rise to multiple sites for protein-protein interactions on the outside perimeter and the inner channel [99, 134]. In addition, the central part of the barrel provides a HC that can recognise methyllysine, eg in EED [134] and WDR5 [138] (Figure 4.11). Additional N-terminal structural elements add to this core structure and provide additional interaction spots. The HC in WD40 proteins is completely different from those found in the RF superfamily and is also less aromatic. In WD40, $\beta 4$ - $\beta 1$ and the $\beta 2$ - $\beta 3$ loops of the β -sheet repeats provide the aromatic and hydrophobic residues of the HC, which is located at the centre of the barrel. This type of cage is formed only on one of the two faces of the barrel. This configuration induces a conformational change of the barrel on peptide binding, which brings all the chains involved in the interaction closer to each other thereby closing the barrel at the binding site.

4.4 Discussion

In this chapter, we characterize the peculiarities of proteins forming the Royal Family. We have also explored the PDB in search of members of this superfamily, identifying cases where the classification in the PDB structure title does not correspond with the classification found on domain databases as pfam, SCOP, interpro or CDD. This is probably due to the fact that the criteria that each database uses to classify a protein differ. For example, the solution structure of SMN Tudor domain (4A4E) is classified as SMN in pfam, which refers to the protein belonging to the SMN family and not to the specific Tudor domain. SMN is a protein family that contain Tudor domains, so the domain is indeed a Tudor domain. Other divergences between databases are related to the independent discovery of the same domain in different proteins. This resulted in different names

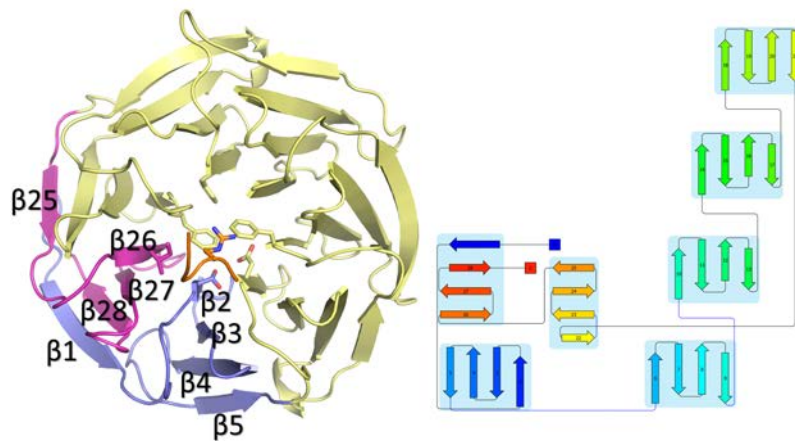


Figure 4.11: Cartoon representation of the structure of human WDR5 in complex with Histone H3K4me2 peptide. Left panel, the structure of the protein is shown as cartoon in pale yellow. The first repeat is coloured in blue, while the last is coloured pink. For clarity, only the β -strands of the first and last repeats are annotated. The HC residues anchoring the dimethyllysine histone peptide are shown as sticks. The Histone H3 peptide is coloured in orange, showing as sticks the dimethylated lysine (PDBid: 2H13 [138]). Right panel, 2D layout of the Human WDR5 structure

for structurally equivalent domains, which then influenced the classification of new structures of similar domains into separate but structurally equivalent families. For example, in our analysis we cannot find differences in the structures of the Chromo Barrel and Tudor-knot domains that justify a separation into different structural families. However, functional differences should be taken into account as well. For example, the RNA-binding activity found in Tudor-knot has been established for some but not all Chromo Barrel domains. Thus, it makes sense to classify Tudor-knot domains as a Chromo Barrel member with RNA binding activity.

Another example of structurally similar domains is the Agenet and Tandem Tudor domains. The initial classification of Agenet domains was based solely on sequence similarity to Tudor domains. The equivalence of the Agenet domains with the independently discovered tandem Tudor domains could not have been detected until the first structures of Agenet proteins were solved. Now that structures of both domains are available, we have been able to show that both consist of a combination of two Tudor domains in tandem and that they are structurally very similar. For this reason, we believe that these domains should be classified as a single subfamily of the Tudor-like family. We propose “tandem Agenet domains” as a name for this combined group, in order to maintain the “Royal Family” essence of this protein superfamily.

Finally, the Chromo Shadow domain has a similar structure to Chromodomains, but the HC and the methyllysine binding function are absent, which would discard this domain as part of the RF. We found in literature that Chromo Shadow domains always coexist with Chromodomains, giving them the ability to self-dimerize. For this reason, we believe that the combination of a Chromodomain and a Chromo Shadow domain should be classified as a single domain, in a similar fashion as in extended Tudor Domains. We propose the name “**extended Chromodomain**” to the combination of a Chromodomain and a Chromo Shadow domain.

The case of the Chromo Shadow domain described above illustrates that classification does not always take into account the combination of domains, organized on a higher level, by separately assigning each domain to a family. However, the coexistence of multiple domains can be important

for function, which in our opinion should be considered. For extended Tudor domains, were the Tudor domain present in the structure is only active if it coexists with the SN-like domain and the C-terminal α -helix, this has been the case. Automatic structure notation of these particular domains struggles with the number of repeats and the high structural similarity, resulting in an overrepresentation of single Tudor domains which are in fact parts of functional modules involving multiples domains. Hybrid Tudor domains are an example of a particular difficult case is in this respect.

In an attempt to provide a consistent classification of RF members that takes into account the considerations discussed above, we summarize below the elements that characterize each of the RF family and subfamilies. First, we summarize the requisites that define the RF superfamily: An RF domain has a typical length of 60-80 amino acids and contains of a hydrophobic cage with the ability to interact with peptides which usually contain arginine or lysine modifications. Furthermore, this hydrophobic cage is located in an antiparallel β sheet composed of three to five β strands. Four different RF families can be distinguished according to their sequence similarity: the Chromo-like and Tudor-like families, the MBT domain and the PWWP domain.

The Chromo-like family consists of three different subfamilies: Chromodomains, Chromo Barrel domains and extended Chromodomains. Chromodomains are the most simple form of RF as they only consist of a three-stranded antiparallel β sheet, and a C-terminal α -helix that packs against it. Chromo Barrel Domains, though, consist of a five-stranded antiparallel β sheet, a $\beta 4$ - $\beta 5$ 3_{10} helix and an optional but frequent $\beta 3$ - $\beta 4$ 3_{10} helix. One of the two additional strands in Chromo Barrel domains, not present in Chromodomains, is located N-terminal to Chromodomain $\beta 1$, while the second is located C-terminal to $\beta 3$. We propose to name the last subfamily of Chromo-like domains as the “extended Chromodomain”, which are characterized by the presence of the Chromo Shadow domain in combination with a chromodomain. The Chromo Shadow domain alone can in our opinion not be considered as a separate RF family as its presence is always linked with Chromodomains, does not present a functional HC, and provides chromodomains with the ability to self-dimerize.

The Tudor-like family consists of four subfamilies: the Single Tudor, the Tandem Agenet, the Hybrid and the extended Tudor domains. The single Tudor domain is the basic structure found in all other Tudor-like subfamilies. In single Tudor domains, a four or five stranded β barrel-like structure harbours a hydrophobic cage, which is the largest HC among all the RF members, consisting of four to five hydrophobic residues, mainly aromatics, and one or two polar amino acids. A 3_{10} helix between $\beta 4$ - $\beta 5$ is present, covering and closing the barrel on one side. Tandem Agenet domains consist of at least two tandem single Tudor domains in a polypeptidic chain, conferring the protein the ability to recognize modifications in more than one residue of the peptide at the same time. Hybrid Tudor domains may be the most dissimilar subfamily within the Tudor-like family, as it consists of two Tudor domains that are interdigitated, sharing $\beta 3$ and $\beta 6$. Finally, extended Tudor domains are the combination of a single Tudor domain and a SN-like domain, complemented with a C-terminal α -helix that link and tighten both domains.

The MBT family members consist of more than two repeats of the MBT fold, which is similar to the Tudor fold, but does not contain the 3_{10} helix between $\beta 4$ - $\beta 5$. The C-terminal sequence forms variable structural elements and interacts with the other MBT repeats of the folded polypeptide, which helps stabilize the final fold. This is the case of the C-terminal α -helix on 2PQW, which interacts with the first MBT and establishes a three-lobulated circular structure.

Finally, the PWWP domain core is formed by a five stranded β -barrel core, which contains an

additional insertion between the second and third strands of the β -barrel and a C-terminal α -helical bundle. The defining characteristic of this domain is the PWWP motif, located in the $\beta 1$ - $\beta 2$ loop. This motif may slightly differ among PWWP domains, but in all cases functions as a nexus between HC and the C-terminal helical bundle. Structural comparison of PWWP domains to other Royal Family members shows that PWWP domain is closer to Tudor and MBT, as all three rely on a five stranded β -barrel core for methylated residue recognition. In contrast, Chromo-like domains, with the exception of Chromo Barrel domains, possess a smaller core composed by of 3 or 4 β strands. The α -helices of PWWP domains are in similar position to that found in MBT domains. The structure of PWWP domains is very similar to Chromo Barrel domains although Chromo Barrel domains do not have the PWWP motif, and PWWP domains lack the 3_{10} helix between $\beta 3$ - $\beta 4$ typical of Chromo Barrel domains [125].

In summary, we have reviewed the main characteristics of each Royal Family domain. We provide characteristics of the different families in order to unify and facilitate the classification of RF structures. We hope that this work will aid in the consistent classification of newly discovered RF-like proteins. As we will cover in the next chapter, ARF-ADs are surprisingly similar to RF members, so this classification will be used in determining the appropriate classification of ARF-AD.



5. In search of an Ancillary Domain function

5.1 Abstract

The ARF Ancillary Domain was discovered in the first *Arabidopsis thaliana* ARF-DBD structures as an independently folded subdomain. It comprises 80 amino acids located in the DBD C-Terminus and is the last structured region before the Middle Region. The fold of this domain strongly resembles that of members of the Tudor Domain, a protein family known for its ability to recognise methylated histone tails. These domains contain a Hydrophobic Cage responsible of the binding and recognition of the methylated peptides, which was not recognized in the ARF Ancillary Domain. Despite this difference, the overall structure of the domain remains highly conserved, suggesting an evolutionary relationship. In this chapter, we carefully analyse the structural similarities and differences of the Ancillary Domains of the ARFs solved to date. This analysis resulted in the identification of a putative hydrophobic cage in a proper conformation for methylated peptide recognition, covered by a flexible loop that hampers access to the HC. We explore the possibility of an HC regulatory function of this loop, which would require activation for recognition. The discovery of a functional reading module on ARFs would completely reformulate the present knowledge of ARF regulation, suggesting that ARF-AD specificity plays a role as well in effective ARF gene regulation.

5.2 Introduction

The structure of ARF-DBDs contains three subdomains. For two of these subdomains, the B3 and dimerization domain (DD), defined functions have been assigned. The function of the last subdomain, the Ancillary Domain (AD), is still unknown, but the structural similarity of ARF-ADs of different organisms suggests it may have a conserved function. The structures of AtARF1 and AtARF5 revealed that the Ancillary Domain was found to be similar to the Tudor domain, a member of the Royal Family [39]. The Royal Family members are histone readers and are responsible of

recognizing different histone posttranslational modifications (PTMs), also referred to as histone marks. Histone tails are extensive PTM hotspots, which are modified by a network of epigenetic reader proteins and enzymes that dynamically add and remove histone marks, ultimately regulating gene transcription through histone mark readers [95]. Editing of the histone marks can turn on and off genes in response to environmental and cellular changes, without altering the underlying genetic information. These changes are the main drivers of cell differentiation, which depends on control of transcription mediated by the histone modifications [84]. The presence of a module with the ability to read these changes would allow ARFs to tune gene expression of the controlled genes in response to the cellular status, integrating epigenetic information into their response. The remarkable structural similarity of ARF-ADs to Tudor domains suggests this histone PTM reading function exists in the ARF-ADs, but experimental evidence for this is lacking. The typical hydrophobic cage found in Royal Family members, a requisite for methyllysine and methylarginine binding, was not evident in ARF-ADs.

In this chapter we reanalysed the ARF-AD in search of a possible function, revealing bioinformatic and structural evidences of a conserved and functional histone PTM reader module within the ARF-ADs. We found that the structural elements observed in ARF-ADs are similar to those found in different Royal family members. Strikingly, the typical Royal Family HC is present, but it is covered by a positively charged amino acid (Arginine/Lysine/Asparagine) of a flexible loop. In fact, the charge and the hydrophobic character at critical positions of the HC are conserved on ARF ADs. Interestingly, there are amino acid sequence variations in the HC composition and surrounding loops that suggest a point for variable recognition of PTM. The Royal Family study presented in chapter 4 serves as the foundation for a thorough structural analysis of the ARF-AD to classify it and understand the biological implication of the conservation of the Ancillary Domain through ARF evolution. Overall, all the necessary elements for a functional Royal Family-like Ancillary Domain are present, but the peculiarities found in ARF-AD differentiate them from the known Royal Family members. All the bioinformatics evidences presented in this chapter are the basis for the design of the experiments presented in chapter 6, where the obtained *in vitro* results demonstrate that AtARF1-DBD is able to bind modified histone peptides.

5.3 Results

5.3.1 ARF Ancillary Domain, founding member of a plant-specific Royal Family-like domain

As was previously reported [39], Dali searches using the Ancillary Domain [39, 114, 117, 139, 140] showed that the Tudor domain 2 of human PHD finger protein 20 (Tud2PHF20, PDBid: 3QII [116]) has a similar structure as the Ancillary Domain (3QII vs AtARF1-AD all atom RMSD: 2.0551 Å). Surprisingly, the 3QII structure also superposed well with the Dimerization Domain of AtARF1-DBD (3QII vs AtARF1-DD all atom RMSD: 2.1917 Å) (Figure 5.1). This suggests that the main structural elements in Tud2PHF20 are conserved in both the DD and AD.

Sequence alignments of the Tudor-like regions found on the AD and DD show that they do not share sequence homology with representatives of Royal Family members included in the databases pfam [141, 142] (pfam clan Tudor (CL0049)) interpro [143] and HMMER [144]. In addition, BLAST searches of the AtARF1-AD sequence for proteins with 30-85% similarity only retrieved ARFs from different plant species but did not result in hits when excluding plants. This indicates that the sequences giving rise to the ARF-Ancillary Domains are very specific of ARF proteins,

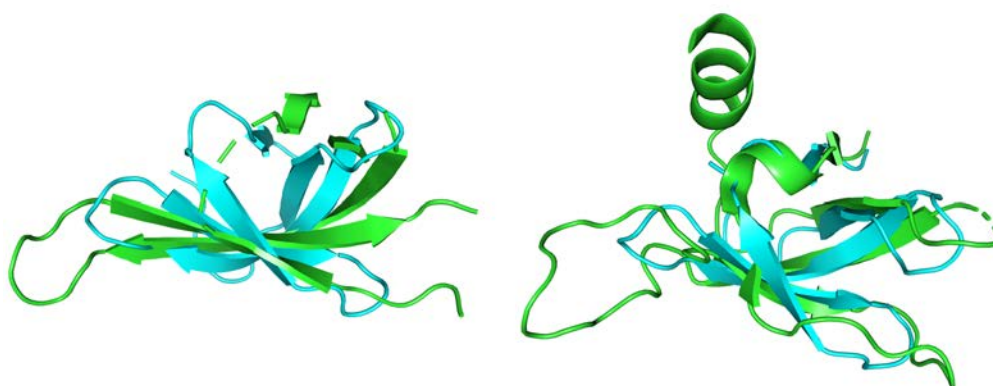


Figure 5.1: Structural superposition of AtARF1AD and DD with 3QII. Left panel: AtARF1DD superposed on 3QII. The AtARF1 $\alpha 2$ and $\alpha 6$ helices were removed for clarity. Right panel: AtARF1AD superposed on 3QII.

despite sharing significant structural similarity with Royal Family Domains. As we cannot rely on sequence homology to compare ARF-AD with other Tudor domains, we embarked on an in-depth analysis of the ARF Tudor-like regions found on DD and AD in order to better understand the characteristics of Ancillary Domains and to establish the possible evolutionary relationship with Royal Family domains.

5.3.2 ARF Tudor-like Dimerization Domain

Structural superposition of Tud2PHF20 with the Dimerization Domain of AtARF1 (4LDX) resulted in an all atom RMSD of 2.1917 Å (Figure 5.2). Although the RMSD value indicates an overall similar structure, significant displacements of all β -strands were observed, which complicated the similarity assessment. For this reason, we analysed the structure of this subdomain using PDBeFold, submitting the coordinates of 4LDX corresponding to the Tudor-like Dimerization Domain in chain A, residues R40-P99 and S238-S276 (B3 and AD subdomains were excluded). The highest scoring match was a Chromodomain, PDBid: 5JJZ, with an RMSD value of 2.08 Å. The resulting structural superposition of all β -strands of the Chromo domain 5JJZ with those of the 4LDX Dimerization Domain resulted in a much better fit of the full structure (Figure 5.2A). The eFOLD search also revealed that the DD motif resembles other Royal Family domains containing 5 antiparallel β -strands. These β -strands fold in a barrel-like structure open on one side, which is reminiscent of the original description of this ARF fold to resemble a “Taco” or a half-closed hand [39]. The β -sheet comprises β -strands 1, 2, 3, 11 and 12 of the AtARF1-DBD, with which two α -helices ($\alpha 2$ and $\alpha 6$) interact, covering the “Taco” fold from the top (Figure 5.2B).

The structural superposition of 5JJZ and AtARF1 DD suggested that the AtARF1-DD hydrophobic cage is incomplete, as only two of the AtARF1 aromatic sidechains, Y43 and F260, shared position with the three hydrophobic cage residues of 5JJZ (Figure 5.2C). It is worth highlighting that the $\alpha 6$ -helix of AtARF1-DD is packed against these residues at a similar localization as the bound peptide in 5JJZ, suggesting that the function of the Royal Family-like domain found in DD could be related to locking this α -helix into place. Given the relevance of ARF dimerization through $\alpha 6$ for ARF function, we do not expect that this α -helix would change position to open the binding pocket of the DD Tudor-like domain. Another structural peculiarity of AtARF1-DD structure is that the $\alpha 2$ in AtARF1 is inserted in the loop connecting $\beta 1$ and $\beta 2$, a known insertion

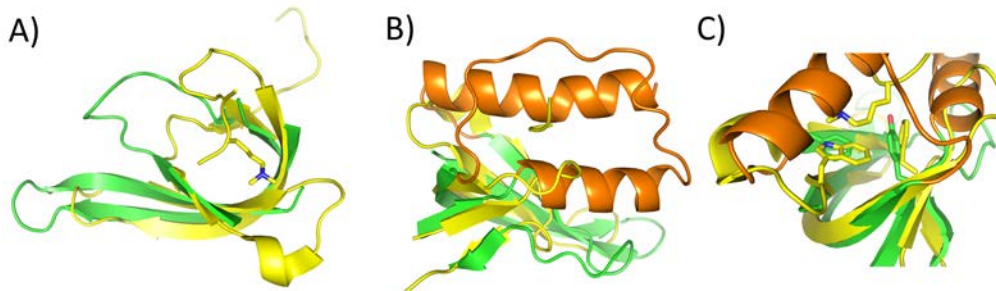


Figure 5.2: Structural comparison of AtARF1 DD (green), and 5JJZ (yellow). A) Superposition of the AtARF1-DD core and 5JJZ. AtARF1-DD α -helices that are not present in 5JJZ were removed for clarity. The sidechain of the histone peptide interacting with 5JJZ HC is shown as sticks B) superposition of the complete AtARF1DD and 5JJZ. The α -helices removed in A are now shown in orange. C) Frontal view of 5JJZ HC. Residues that structurally align with AtARF1DD are shown as sticks. The α_6 and α_2 helices covering the DD HC are coloured orange.

point of varying secondary structure elements in Royal Family members. In contrast, the B3 domain and α_6 are inserted in the loop connecting β_3 and β_4 . In Royal Family members insertions at this point have not been observed [125, 140, 145].

5.3.3 ARF-DBD Tudor-like Ancillary Domain

All the ARF-DBD structures solved to date include Ancillary Domains in the last 80 C-terminal amino acids, which adopts an open barrel structure with three antiparallel β -strands, whereas the equivalent region to the fourth β strand of RF proteins is much shorter and is a random coil. Similar to what was observed for Royal Family members, two loops of variable length connect β_1 to β_2 and β_2 to β_3 [125, 140, 145]. A C-terminal 3_{10} helix is present in the loop connecting the small β_4 - β_5 strands, which comprises residues 348-352 (AtARF1 numbering) and lies above the open part of the barrel (Figure 5.3). This 3_{10} helix is characteristic of the Ancillary Domains and is also found in Chromo- and Tudor-like domains [99, 101]. A structural superposition of AtARF1-AD with the Tud2PHF20 (3QII) protein presented above shows that the β -strands of the Tudor domain containing the hydrophobic cage of Tud2PHF20 have similar counterparts in the AtARF1 Ancillary Domain. Further analysis localized five putative amino acids in the AtARF1 Ancillary Domain that structurally align with the Tud2PHF20 HC residues, and therefore could form an ARF Ancillary Domain HC (Figure 5.3). 3QII HC residues W97, Y103, F120, D122 and V124 share position with AtARF1-DBD F298, F308, W335, E337 and F342, respectively. Although the residues are not identical, the overall hydrophobic and charge contribution are preserved, indicating putative PTM recognition in ARF-ADs (Figure 5.3, Right panel). Despite the similarity, the main difference between the Royal Family and AtARF1-AD structures is in the loop connecting β_3 and the 3_{10} helix, which is a random coil in AtARF1, whereas this regions folds as a β_4 -strand in almost all Royal Family members (Figure 5.3).

The above analysis suggests the AD as a putative histone reader. In light of the possible inclusion of a functional Royal family domain in ARFs, we compared ARF-AD HCs with several Royal Family HCs in complex with ligands and performed structural superposition of the proposed HC residues between ARFs from *Arabidopsis thaliana* and distant relatives from *Marchantia polymorpha* and *Chlorokybus atmophyticus* [55].

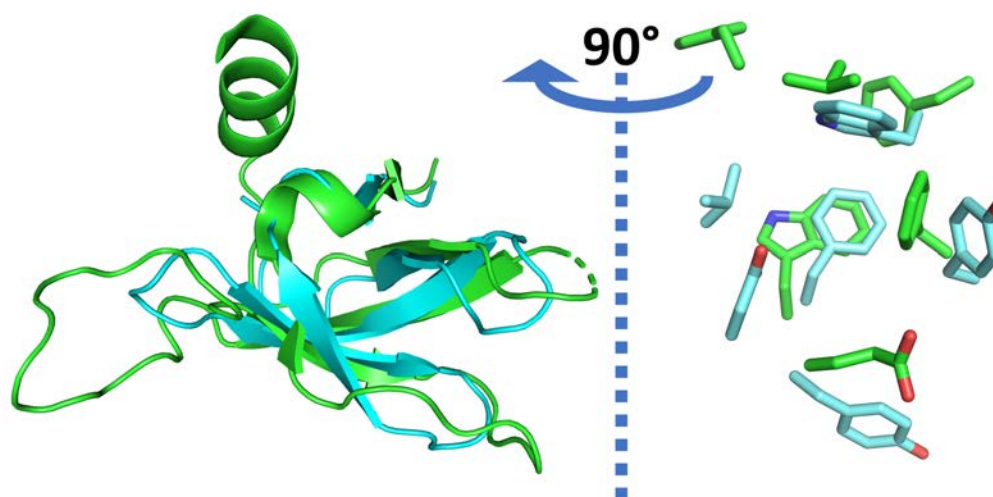


Figure 5.3: Structural superposition of AtARF1 AD with Tud2PHF20. Left panel: Cartoon representation of the side view of the structure of the second Tudor domain of human PHD finger protein 20 (3QII, blue), superposed on the structure of the AtARF1 AD (4LDX, residues 277-356, green). The HC residues in AtARF1 (green sticks) were assigned by structural superposition with HC residues of Tud2PHF20 (blue sticks). Right panel: Frontal view of the HC residues of AtARF1 (green sticks) and Tud2PHF20 (blue sticks). The cartoon representation was removed for clarity.

5.3.4 The ARF-AD Hydrophobic Cage is structurally related to the Royal Family HC

The structural superposition of AtARF1 AD with the structure of Tud2PHF20 suggested the presence of the HC in AtARF1 formed by five candidate residues. We superposed ADs of the ARF-DBDs present in the Protein Data Bank to check for the structural conservation of the HC. The structures of *Arabidopsis Thaliana* ARF1, ARF5 and *Marchantia polymorpha* ARF2 ADs all displayed residues compatible with functional HCs (Figure 5.4). Of special relevance is the complete structural conservation of W335/370/339 and E337/372/341 in AtARF1, AtARF5 and MpARF2 structures, respectively. Two additional aromatic residues in AtARF1 (F298 and F308) have similar counterparts in AtARF5 (F333 and Y343) and in MpARF2 (F302 and H312). On the other hand, F342 in AtARF1 is the least preserved of the predicted HC amino acids, as AtARF5 and MpARF2 contain charged residues, D377 and E346 respectively, in this position. It should be noted that the sidechain conformation of the conserved residues is identical in different structures of the same ARF, that the electron density was unequivocal and that none of them showed alternate conformations.

In order to compare the HCs of ARF structures to that of the RF proteins, Figure 5.5 shows the HC regions of representative members of each of the Royal Family subfamilies in the same orientation as those shown in Figure 5.4 for the ARFs. A minimum of three aromatic residues are present in these HCs, which is the case for structures 1KNA, 2QQS and 5CIU. Additional aromatic residues are found to interact with the methyl group of the bound methylated peptide in the other RF structures shown in Figure 5.5. Furthermore, non-aromatic residues may contribute to the specificity of the interaction. These residues can be acidic (2LVM, 2QQS, 2PQW, 5CIU) or polar (3OMC, 2QQS) and interact with the amino group of lysine. This is similar to what is observed in the ARF-ADs structures, where at least three aromatic residues are present, and the remaining two are acidic or aromatic. It can be seen in Figure 5.5 that the position of the peptide bound is different in the Royal Family members. As not all RF members share the same secondary structure

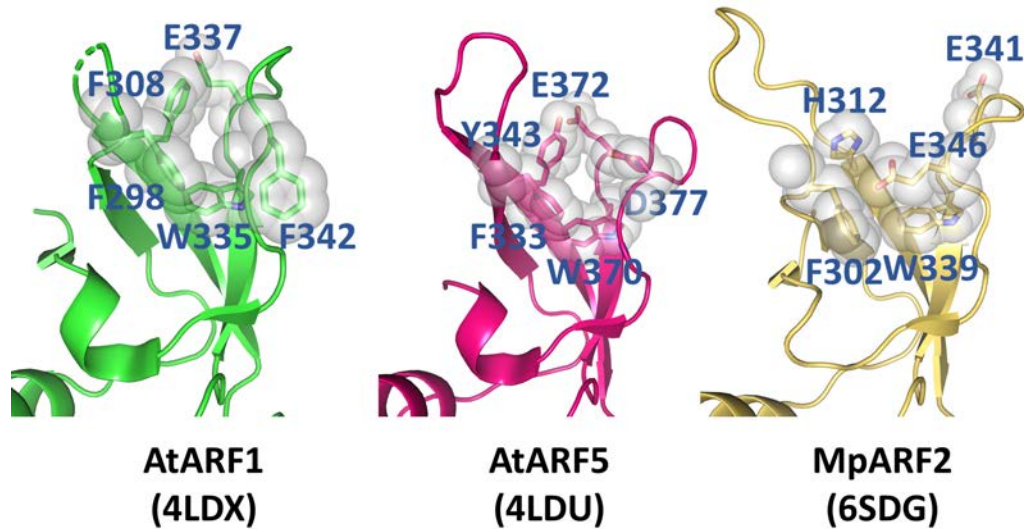


Figure 5.4: Structural comparison between known ARF-ADs. The putative HC residues are shown as sticks, with an overlaid sphere representation to delineate the HC pocket. PDB codes are shown in parenthesis.

elements, the opening of the HC may differ, resulting in different orientations of the bound lysine and/or arginine residues.

5.3.5 Putative ARF cage residues are conserved during evolution

The identification of residues of the HC of the AtARF1-AD from the structural alignment with Tud2PHF20 was the starting point to look for conservation of these residues in other ARFs. We included in the analysis all *Arabidopsis thaliana* ARFs, *Marchantia polymorpha* ARFs (MpARFs) and a recently identified ARF in the green algae specie *Chlorokybus atmophyticus* (CaARF) [55]. MpARFs and CaARF were of interest because of they represent early occurrences in the evolution of the ARFs and therefore can be used to probe the evolutionary conservation of the HC in the AD. Our results show that the equivalent positions to AtARF1 F298, F308, W335, E337 and F342 are highly conserved during evolution (Table 5.1). Remarkably, W335 is completely conserved and F298 and E337 are highly conserved. In addition, F308 is either a phenylalanine or a tyrosine in all the studied ARFs. Interestingly, residue F342 is substituted by charged residues in class A ARFs, which suggests that these may interact with a positively charged ligand. This residue is a glutamine in all *Arabidopsis* and *Marchantia* class C ARFs and this remarkably conservation suggests that this domain was present before class A/B and C ARFs diverged. Together, these observations reveal an evolution-driven conservation of these residues, which suggest that the ARF HC may be involved in a function.

An interesting observation is the contrast in conservation of individual residues in the different classes ((Table 5.1)). For example, the residues, in particular the aspartate and glutamates that are equivalents of AtARF1 F342, are much conserved in Class A than in Class B, which is also clear from the sequence signatures. These residues also show high conservation in Class C ARFs but there are class B ARFs (ARF11 and ARF18) which have a signature that coincides with the Class C ARFs, since residues 308 and 342 are F and Q respectively. However, overall, the signatures match well with the classification of the ARFs, which is based on the full sequence. This suggests that specificity of binding for the recognition partner of the AD is similar within each class.

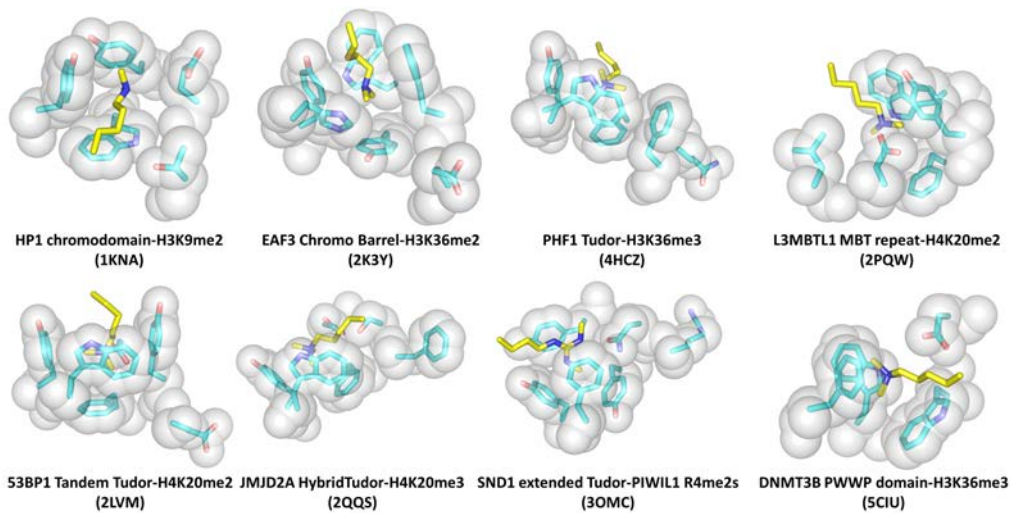


Figure 5.5: Hydrophobic Cages from different Royal Family members in complex with methylated residues. The Hydrophobic Cage residues are coloured in blue and shown as sticks with overlaid sphere representation. The interacting methylated residues are shown as yellow sticks. The PDB code of each structure is shown in parenthesis. The images shown in this figure were obtained after a structural superposition of all the structures, maintaining the same camera angle for the snapshot. Differences in orientation of HC or binding residues are in fact variations in the orientation of the binding cage between Royal Family domains.

Table 5.1: Conservation of AtARF1 residues F298, F308, W335, E337 and F342 in *Arabidopsis thaliana*, *Marchantia polymorpha* and *Chlorokybus atmophyticus* ARFs

| Class | ARF | F298 | F308 | W335 | E337 | F342 | Signature | Class | ARF | F298 | F308 | W335 | E337 | F342 | Signature |
|-------|-------|------|------|------|------|------|-----------|-------|-----|------|------|------|-------|-------|-----------|
| A | 5 | F | Y | W | E | D | FYWED | 1 | F | F | W | E | F | FFWEF | |
| | 6 | F | Y | W | E | E | FYWEE | 2 | F | F | W | E | P | FFWEP | |
| | 7 | F | Y | W | E | D | FYWED | 3 | V | S | W | D | G | VSWDG | |
| | 8 | F | Y | W | E | E | FYWEE | 4 | F | C | W | E | D | FCWED | |
| | 19 | F | Y | W | E | D | FYWED | 9 | F | Y | W | E | S | FYWES | |
| | Mp1 | F | Y | W | E | E | FYWEE | 11 | F | F | W | E | Q | FFWEQ | |
| C | 10 | F | F | W | E | Q | FFWEQ | 12 | L | C | W | E | P | LCWEP | |
| | 16 | F | F | W | E | Q | FFWEQ | 13 | F | Y | W | E | L | FYWEL | |
| | 17 | M | F | W | E | Q | MFWEQ | 14 | F | S | W | E | P | FSWEP | |
| | Mp3 | F | F | W | E | Q | FFWEQ | 15 | F | Y | W | E | L | FYWEL | |
| | CaARF | C | C | W | D | H | CCWDH | 18 | F | F | W | E | Q | FFWEQ | |
| | | | | | | | | 20 | F | Y | W | E | S | FYWES | |
| | | | | | | | 21 | F | Y | W | E | S | FYWES | | |
| | | | | | | | 22 | F | Y | W | E | S | FYWES | | |
| | | | | | | | 23 | - | - | - | - | - | - | | |
| | | | | | | | Mp2 | F | H | W | E | E | FHWEE | | |

5.3.6 Sequence homology of the ARF ancillary domain

We have shown above that the HC residues are highly conserved, but this could be the result of high conservation of the whole domain or because these residues are important for domain folding. To visualize the overall conservation of the AD and that of selected residues, we computed a logo [146] based on sequence alignments of the *Arabidopsis thaliana*, *Marchantia polymorpha* and *Chlorokybus atmophyticus* sequences performed with Clustal omega [147] (Figure 5.6). The computed logo show patches of high conservation, for example in the WDE 335-337 motif or the C-terminal residues 347-353. The alignment also highlights the complete conservation of all the tryptophan residues in the sequence (323,328,335 and 350), and shows the high variability in N-terminal region (residues 277-290) and in the loop C-terminal to $\beta 3$ (residues 338-342).

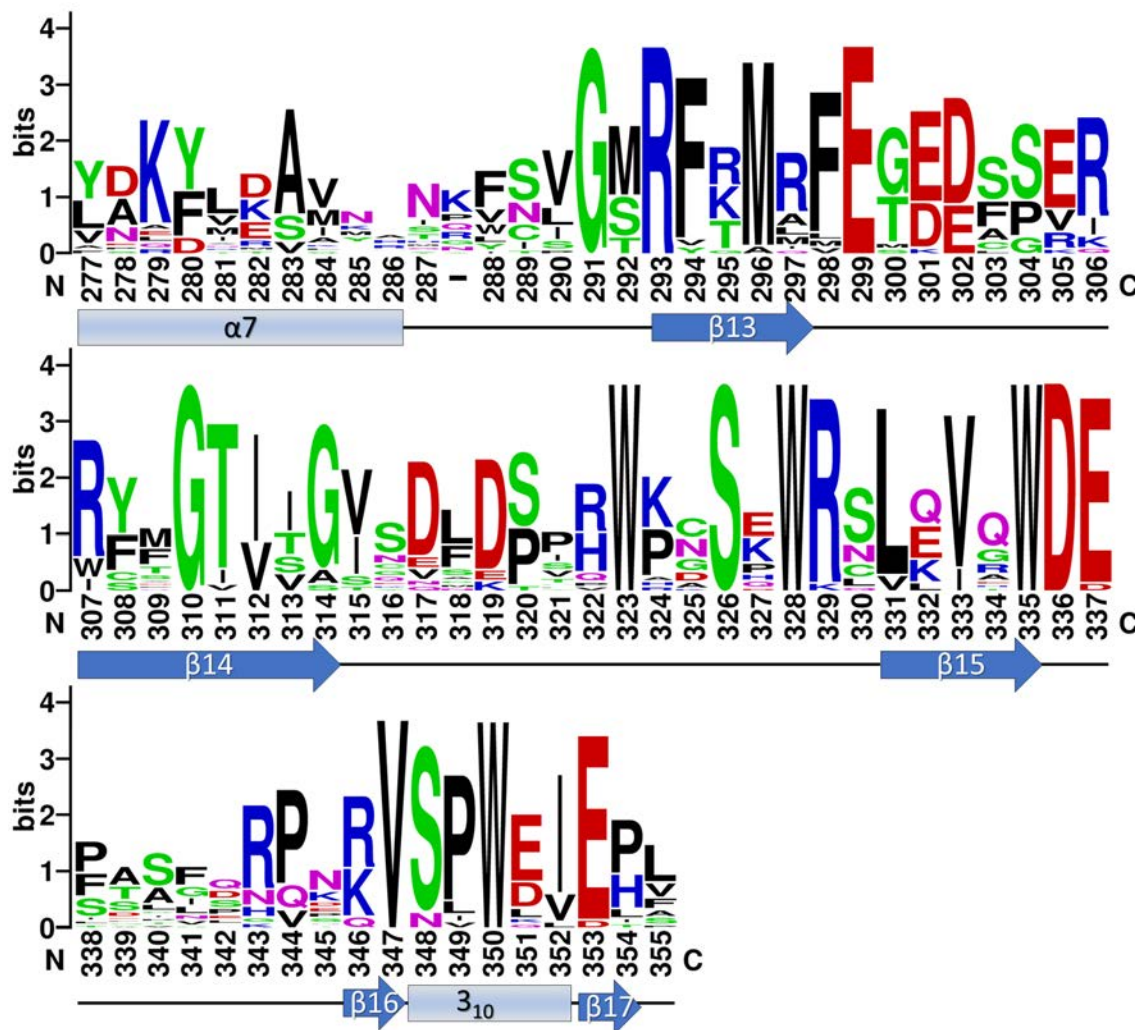


Figure 5.6: Weblogo of the sequence alignment of all the *Arabidopsis* ARF-ADs, aligned with Clustal Omega to AtARF1AD. Numbering corresponds to AtARF1. Amino acids are coloured according to their chemical properties: polar amino acids (G,S,T,Y,C,Q,N) in green, basic (K,R,H) in blue, acidic (D,E) in red and hydrophobic (A,V,L,I,P,W,F,M) amino acids in black

To visualize the presence and variability of the Ancillary Domain in other proteins and species, we used the webserver ConSurf [148, 149, 150, 151, 152]. For this analysis, we submitted the structure of the AtARF1 AD (residues 277-355, PDBid: 4LDX [39]) to the ConSurf algorithm

using the default options. The server retrieved 300 sequences with homology ranging from 35 to 95%, of which 292 were annotated as Auxin Response Factors, 2 as transcription factors (Uniprot S1SI39 and B9S3X2) and the remaining 8 were annotated as uncharacterized.

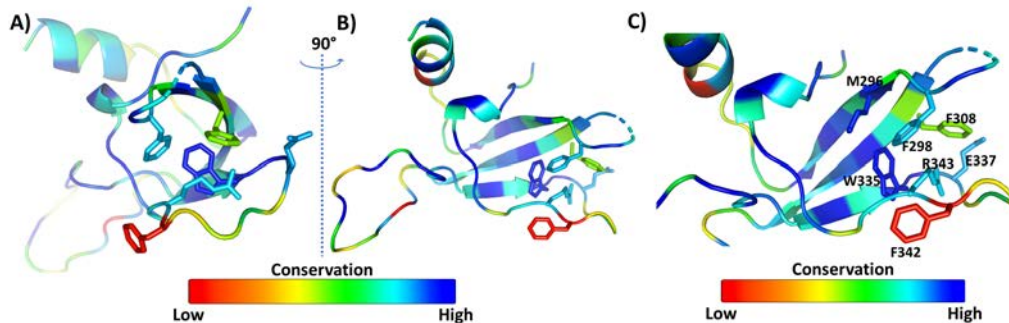


Figure 5.7: ConSurf results of AtARF1 Ancillary Domain conservation. Residues are coloured based on their sequence conservation as calculated by ConSurf. The residues suggested by structural alignment with Tudor domains as the putative Hydrophobic Cluster are shown as sticks. A: Front view of the Ancillary Domain on the entrance to the hypothetical HC is located. B: Top view of the Ancillary Domain generated by rotating the orientation shown in panel A by 90 degrees. C: Detail of the AtARF1-AD HC, where the residues detected by structural alignments as part of the HC are labelled.

The ConSurf results (Figure 5.7) facilitate the interpretation of the conservation suggested by the sequence alignments of *Arabidopsis thaliana*, *Marchantia polymorpha* and *Chlorokybus atmophyticus* ARFs. Thus, W335 is highly conserved, and the degree of conservation of the other predicted HC important residues is also reasonably high. The exception is the F342, which is marked as a low conservation residue. As is shown in structural representations of ARF-ADs (Figure 5.7), the sidechain of F342 points outwards the HC, which may explain why this residue is highly variable, although its $C\alpha$ position is similar to the corresponding RF HC residue. Thus, F342 may not be essential for HC function (Figure 5.7C).

5.3.7 The Hydrophobic Cage entrance is regulated by a basic residue

Previously we pointed out that the loop connecting the $\beta 3$ strand and the 3_{10} helix (residues 336-346) is a random coil in ARF-ADs. The location of this loop corresponds to that the $\beta 4$ strand in Royal Family members. Interestingly, a residue of this loop, R343 (in AtARF1), is interacting with the main chain of E337, P338 and S339 residues, also located in this loop of the Ancillary Domain. This interaction locks the position of the loop in the observed conformation (Figure 5.8). The residue is not only locking the loop connecting $\beta 3$ and 3_{10} helix into place, it is also occluding the putative binding site of a lysine or arginine (Figure 5.8). The interaction of R343 with the main chain of E337, P338 and S339 residues may prevent the formation of the β strand present in other Tudors. Release of this anchor point is a requisite for a functional Hydrophobic Cage, and a mechanism to control this release would result in a regulation of the putative ARF-AD interaction with a methylated basic residue. Interestingly, S340, positioned just after the residues interacting with R343, is predicted by Netphos [153, 154] to be phosphorylated, probably by the cdc2 kinase, a key regulator of cell cycle (Casein kinase II consensus phosphorylation site: [ST]-x(2)-[DE], most ARFs contain the motif 347-SPWD/E-352). Other serine and threonine residues within this loop are also potential targets of kinases. Thus, these residues represent additional points for possible regulation of the position of the loop.

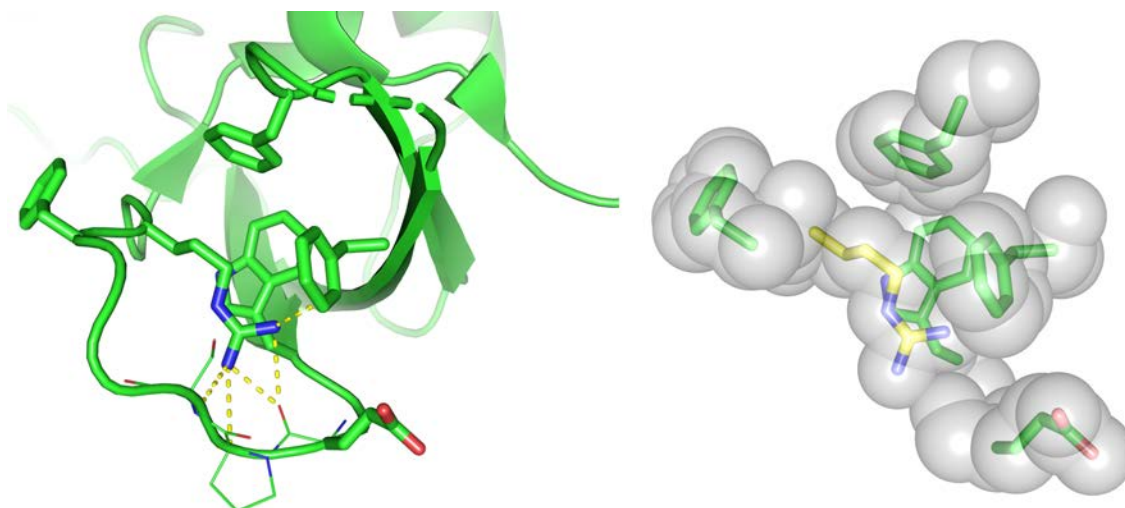


Figure 5.8: Arginine 343 in AtARF1 interacts with b3-b4 loop, blocking the HC entrance. Left panel: AtARF1 R343 interaction with connecting loop β 3- β 4 as observed in (4LDX). Right panel: Sphere representation of R343 (yellow), blocking the entrance of the HC.

Table 5.2: AtARF1 R343 conservation in *Arabidopsis thaliana*, *Marchantia polymorpha* and *Chlorokybus atmophyticus*

| Class A | | | Class B | | | Class C | | | |
|---------|---|----|---------|----|---|---------|---|-------|---|
| 5 | K | 1 | R | 12 | R | 20 | R | 10 | N |
| 6 | R | 2 | R | 13 | R | 21 | G | 16 | N |
| 7 | R | 3 | H | 14 | R | 22 | R | 17 | N |
| 8 | R | 4 | H | 15 | R | 23 | - | Mp3 | G |
| 19 | R | 9 | R | 18 | R | Mp2 | R | CaARF | Q |
| Mp1 | R | 11 | R | | | | | | |

Similar to AtARF1, the structure of AtARF5 and MpARF2 also show an occluded HC (Figure 5.9). In the MpARF2 structures, the density maps are weak at the loop region compared to the residues of the HC, suggesting that this region is intrinsically mobile (Figure 5.9, pink). In fact, compared with the higher resolution MpARF2 structure (Figure 5.9, yellow) it can be seen that the HC binding site may be occupied by two loop residues, E346 and R347 respectively, in each structure. The conservation of this residue in other ARFs (Table 5.2) reveals that this amino acid position is mostly occupied by basic residues. From the analysed ARFs, most possess an arginine (17/27) whereas others an asparagine (3/27, only class C), a histidine (2/27), a lysine (1/27) or a glutamine (1/27) in this position. Interestingly, 2 out of 27 possessed a glycine residue, which under our hypothesis suggest a constitutively active AD (Table 5.2). In contrast to what was observed for the conservation of the HC residues, the conservation of R343 within the three ARF classes matches their classification very well. ARF23 does not have a residue equivalent to R343, as it is truncated in the equivalent position of AtARF1 R223, right after the end of the B3 domain. Consequently, the hypothetical protein would only comprise the first three β strands and two α helices of the DD and a complete B3. For this reason, ARF23 is often marked as a pseudoARF [33, 40].

The presence of a blocking residue in the binding pocket of the hydrophobic cage is not a common feature in Royal Family, but when it is present, it is generally attributed to nonfunctional domains [106, 109, 116]. We found some structural examples of this type of RFs (Figure 5.10).

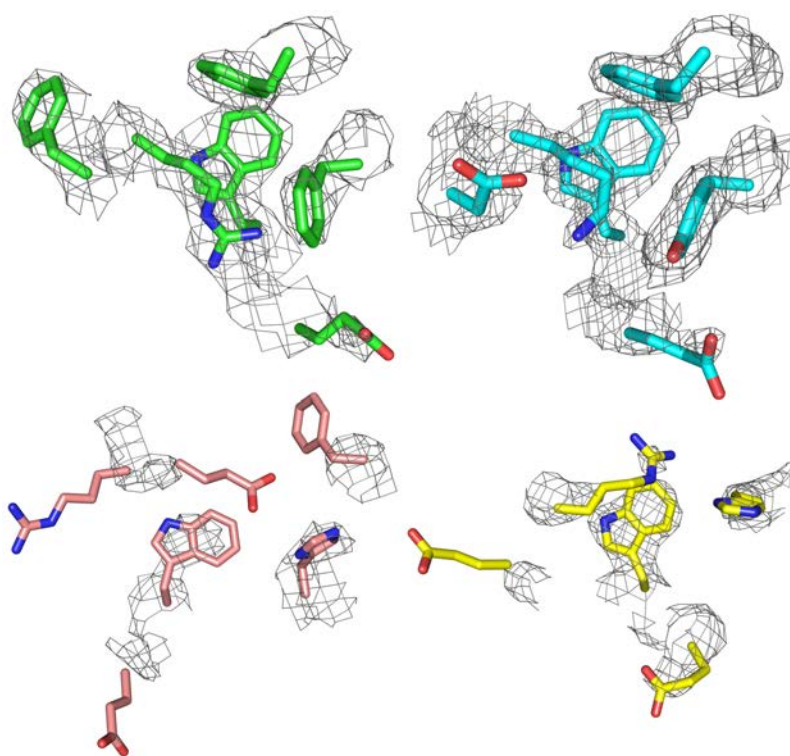


Figure 5.9: Electron density map of HC and interacting residues contoured at 1σ . The AtARF1-4LDX structure is coloured in green, AtARF5-4LDU is coloured in blue, MpARF2-6SDG is coloured in pink, and the structure of MpARF2-ER7 is coloured in yellow. Note that the density in MpARF2-6SDG structure is very low in this region.

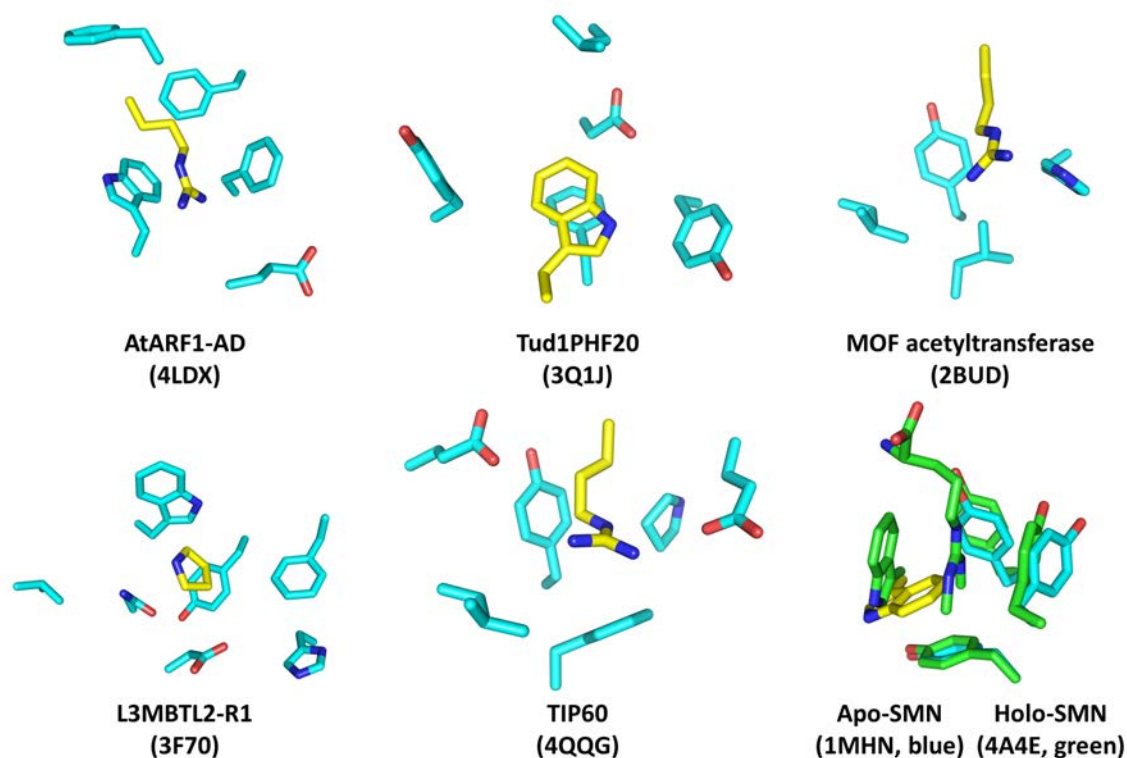


Figure 5.10: Comparison of several Royal Family structures with occluded but properly formed HCs. The amino acid sidechain of the HCs of several occluded royal family members are shown as sticks, while the occluding residue sidechain are shown as yellow sticks.

For example, the Chromo Barrel domain of TIP60 contains a basic residue (R17) whose side chain occupies the cage (Figure 5.10, 4QQG). Despite containing a properly formed hydrophobic cage, R17 prevents this domain to function as methyl amino acid reader [109]. The authors, though, mention that the electron density is not clear and they leave open the possibility of a competitive regulation of cage entrance [109]. Another example is the Chromo Barrel domain of MOF acetyltransferase, in which R387 occupies a position that would clash with a methylated residue entering into the HC (Figure 5.10, 2BUD) [106]. Examples in non-chromo domain Royal Family members are also found. Thus, W50 of the first Tudor domain of human PHD finger protein 20 (Tud1PHF20) is blocking the entrance of the HC [116]. The function of Tud1PHF20 is unknown [99, 116], but the possibility that W50 could move and function as an entrance regulator would explain why Tud1PHF20 contain a properly formed Hydrophobic Cage [116, 155].

Interestingly, in some structures where the HC is blocked by an inhibitory residue, the proteins have been shown to undergo a structural reorganization that opens the HC. An example of this is found in the structure of the SMN Tudor domain, which shows an occluded HC entrance in its Apo form (PDBid: 1MHN, [156]). Interestingly, the HC is opened upon binding of a symmetrically dimethylated arginine [122] (PDBid: 4A4E, [157]). It should be noted that the blocking residue present in SMN is localized in the loop $\beta 1$ - $\beta 2$, in contrast to the locking residue in ARF-ADs, which is localized in the loop $\beta 3$ - $\beta 4$.

In the MBT domain family, an example of a properly formed and occluded HC is found in the L3MBTL2-H4K20me1 structure. This protein is of special interest for our analysis as it displays four repeats of the MBT motif in the structure, where only the fourth MBT repeat (L3MBTL2-R4)

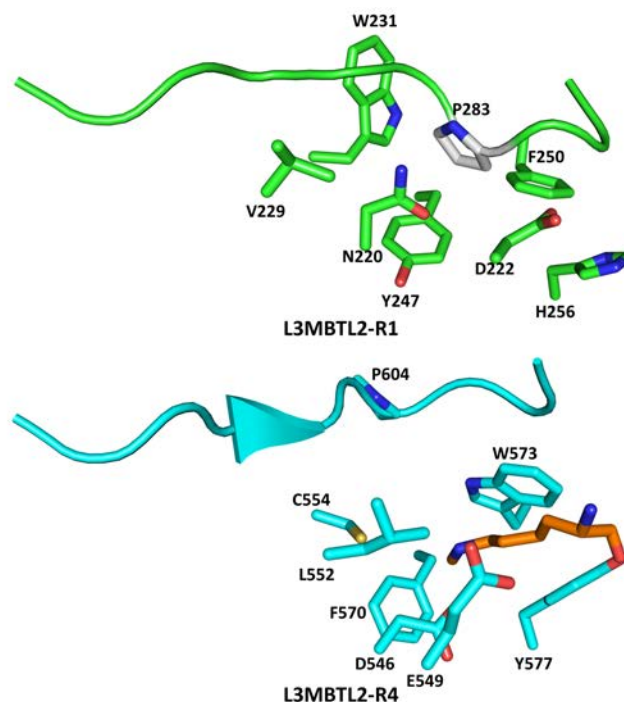


Figure 5.11: Alternate conformation of the loop containing the blocking residue P283 in L3MBTL2 repeat 1 and the open conformation of P604 in L3MBTL2 repeat 4 (3F70). Repeat 1 has a different HC composition than repeat 4, favouring the autointeraction and preventing the binding of the peptide. The loop containing P604 in L3MBTL2-R4 is in open conformation, allowing the interaction.

binds the methylated peptide. Domains two and three lack an HC. The first (L3MBTL2-R1) contains a properly formed HC (Figure 5.11, top panel), but this cage is blocked by residue P283 [99, 158]. Comparison of the loop containing P283 of L3MBTL2-R1 with that of L3MBTL2-R4 reveals that this loop is actually in an alternate conformation in repeat 1 compared to repeat 4, where in the latter the interaction with the assayed peptide is more stable (Figure 5.11, bottom panel). The difference in affinity towards the proline of the loop or the assayed peptide in repeats 1 and 4 can be attributed to a different composition of the hydrophobic cage, which implies different substrate selectivity between repeats 1 and 4.

The structure of the Tudor domain of TDRD3 in complex with a small molecule inhibitor (PDBid: 5YJ8 [159]) further reinforces the hypothesis of a competitive autointeractor entrance regulator in some RF members. TDRD3 is a multidomain protein, containing a Tudor domain comprising residues 555-615. In the structure, an asparagine residue (N596) is blocking the HC, which is in an equivalent position to AtARF1 R343. The authors found an extraordinary large chemical shift perturbation of residue N596 upon addition of the inhibitor in NMR analysis, which they attributed to either a strong interaction with the substrate or to a structural rearrangement of N596 [159]. Several X-ray structures for this protein (Uniprot Q9H7E2) are available containing this region, *i.e.* 5YJ8 [159], 3PMT [160], 3S6W [160], 3PNW [161], as well as an NMR structure (2LTO [162]). We superposed the 3PNW, 5YJ8 and 2LTO structures, which revealed a high mobility of the N596 residue (Figure 5.12), suggesting that a structural change of the β 3- β 4 loop is possible. This could be highly relevant to ARFs, particularly considering that this position is also occupied by asparagine in some ARFs.

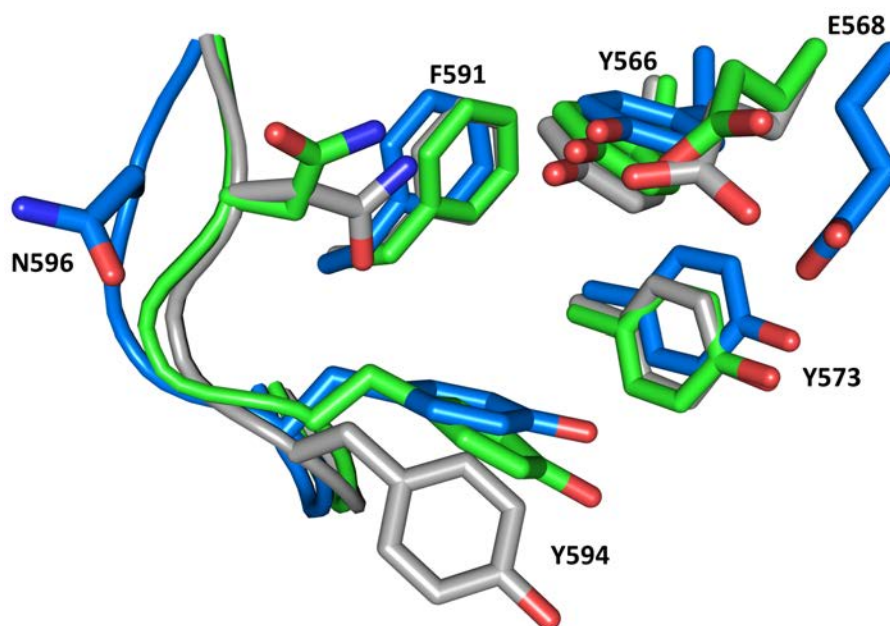


Figure 5.12: Superposition of different structures of the Tudor domain of TDRD3. The structure of TDRD3 in complex with asymmetrically dimethylated arginine (PDBid: 2LTO, in blue) and that of TDRD3 in complex with a small molecule inhibitor (PDBid: 5YJ8, coloured in green) are superposed on the structure of the Apo protein (PDBid: 3PNW, coloured in grey). The sidechains of the HC residues and N596 residue are shown as sticks, while the flexible loop containing the N596 residue is shown as cartoon.

Overall, these results show that cases exist in which a polar or basic residue blocking the HC entrance can move out upon a structural reorganization, which suggests that the blocking residues may function as an affinity regulator. It would be interesting to check whether R343 has a similar function in ARFs.

5.3.8 The Ancillary Domain shares structure with most Royal Family domains

The initial structural alignments of the Ancillary Domain of AtARF1 and AtARF5 with RF proteins showed that the AD belongs to the Tudor domain family [39], but it is not yet clear to which specific RF family the ARF-AD belongs. In order to determine which structural elements are shared between ARF-ADs and RF domains, we performed systematic superpositions of the ARF ADs on representatives of the RF domains discussed in chapter 4. For Tudor domains, the superpositions of AtARF1 and AtARF5 were done manually using Superpose [163] from the CCP4 package [164], selecting several well-known human Tudor domains [101, 114, 155, 162]. Tudor domains required manual superpositioning as the superposing program did not properly superpose hybrid and extended Tudor domains. For the other RF domains, the superposition was done using a script, where multiple structural superpositions were programmatically performed running the CCP4 program Superpose against structures retrieved from the Protein Data Bank. The analysed structures were obtained by searching the structure titles of the entries in the Protein Data Bank for the terms “chromodomain”, “Chromo Barrel”, “Chromo Shadow”, “MBT”, “PHD finger” and “PWWP”. 922 PDBs were obtained. The family assessment of the structures was done by the pfam family name of each PDB entry.

Table 5.3: Structural superpositions of RF domains to AtARF1(4LDX) and AtARF5 (4LDU) AD structure

| PDBid | Subfamily | RMSD (4LDX, Å) | RMSD (4LDV, Å) | F298 | F308 | W335 | E337 | F342 | R343 | Signature |
|-------|---------------|----------------|----------------|------|------|------|------|------|------|-----------|
| 3PNW | Tudor | 1.2303 | 1.5311 | W | Y | F | D | Y | E | WYFDY |
| 3P8D | Tudor | 1.4619 | 1.6628 | W | Y | F | D | V | Q | WYFDV |
| 1G5V | Tudor | 1.3279 | 1.4578 | W | Y | Y | G | Y | N | WYYGY |
| 2LVM | Tudor | 1.3915 | 1.8377 | W | Y | F | D | Y | E | WYFDY |
| 4BD3 | Tudor | 1.3657 | 1.5475 | Y | W | F | E | Y | K | YWFY |
| 4HCZ | Tudor | 1.3927 | 1.5116 | W | Y | F | E | F | Q | WYFQ |
| 4H75 | Tudor | 1.5412 | 1.5193 | F | W | Y | E | Y | L | FWYEL |
| 3ME9 | Tudor | 2.044 | 2.112 | Y | Y | F | D | P | S | YYFDP |
| 2GFA | Tudor | 1.4364 | 1.7968 | W | Y | F | D | F | F | WYFDF |
| 5HH7 | PHD | 1.1067 | 1.2829 | R | W | W | M | N | D | RWWMN |
| 6AT0 | Chromo | 1.2288 | 1.6835 | - | - | W | G | - | E | -WG- |
| 3PMI | PWWP | 1.2739 | 1.5387 | H | W | Y | E | - | K | HWYE- |
| 1PFB | Chromo | 1.3034 | 2.9324 | - | Y | W | G | - | Y | -YWG- |
| 6V2H | Chromo | 1.3529 | 3.142 | A | Y | W | G | - | E | AYWG- |
| 2L1B | Chromo | 1.3591 | 2.8021 | - | F | W | G | - | Y | -FWG- |
| 3KUP | Chromo Shadow | 1.3865 | 2.3003 | F | D | W | D | - | - | FDWD- |
| 2DAQ | PWWP | 1.407 | 1.4114 | L | W | F | G | - | - | LWFG- |
| 3OB9 | Tudor | 1.4128 | 2.134 | E | Y | F | G | - | W | EYFG- |
| 3M9Q | Chromodomain | 1.4348 | 1.4335 | E | Y | F | G | - | Y | EYFG- |

The RMSDs of the backbone atoms and the conservation of the putative HC residues in ARF-ADs and in Tudor domains are shown in Table 5.3. A good agreement between single Tudor domain structures and AtARF1/5-ADs is found, with backbone RMSDs below 2Å in most cases. In addition, the vast majority of Tudor structures contain a tryptophan residue as being part their HC, which is also observed in ARF-ADs. This tryptophan residue, when present in Tudor domains, is important for ligand binding, and in ARFs, it interacts with the arginine blocking the HC (Table section ARFAD). The amino acids found at the positions corresponding to the AtARF1 HC aromatic residues F298, Y308 and W335 were mostly aromatic in the Tudor domain analogues. The glutamate AtARF1 E337, conserved in ARFs, is also conserved in Tudor domains, where it is either an aspartate or a glutamate. AtARF1 F342, which is highly variable in ARFs, is much more conserved in the analysed Tudor domains and alternates mostly between tyrosine and phenylalanine. Finally, the residues corresponding to the putative regulatory R343 in AtARF1 are also charged residues in most of the Tudor domains analysed, with the exception of 2GFA, in which this position is occupied by a hydrophobic residue.

Next, we superposed the Ancillary Domain of AtARF1, AtARF5 and MpARF2 from the PDBs 4LDX, 4LDU and 6SDG, respectively, on the non-Tudor RF structures retrieved from the PDB. Our alignments of ARFs AD with the retrieved RF domains show high divergence in the β 2- β 3 loop, which is in line with findings of other studies on RF members [125, 145]. For example, superposition of the PWWP domains shows that the insertion between the second and third β strands varies in length and in secondary structure among these different classes of PWWP domains. The authors propose that the variability in this region is likely caused by intron/exon sliding at the genomic level, as the coding region for the second and third β strands are often split by an intron [125, 145]. This variable loop is also the insertion point where tandem Tudor repeats share the β strands. The implication of the variability of this loop for the function of the ARFs must be analysed, but for our analyses we manually removed this variable loop in order to improve the superpositions.

The top ten ranked structures in the multiple superpositions with non-Tudor RF domains are included in Table 5.3 and consisted on Chromodomains (5/10) with the lowest RMSD value being

Table 5.4: Amino acid frequency in non-Tudor RF domains based on AtARF1 HC

| Amino acid | F298 | F308 | W335 | E337 | F342 | R343 | Amino acid | F298 | F308 | W335 | E337 | F342 | R343 |
|------------|------|------|------|------|------|------|----------------------|------------|-----------|-----------|------------|-----------|-----------|
| A | 6 | 0 | 0 | 0 | 6 | 0 | N | 0 | 0 | 0 | 2 | 1 | 2 |
| C | 1 | 16 | 0 | 0 | 0 | 0 | P | 0 | 0 | 0 | 2 | 3 | 0 |
| D | 31 | 1 | 0 | 33 | 1 | 11 | Q | 5 | 0 | 1 | 0 | 0 | 1 |
| E | 5 | 0 | 1 | 9 | 4 | 20 | R | 9 | 9 | 7 | 2 | 0 | 5 |
| F | 2 | 18 | 49 | 3 | 0 | 6 | S | 3 | 0 | 0 | 5 | 5 | 1 |
| G | 2 | 7 | 0 | 70 | 6 | 11 | T | 1 | 0 | 0 | 0 | 4 | 9 |
| H | 13 | 1 | 1 | 0 | 0 | 0 | V | 2 | 1 | 6 | 0 | 1 | 0 |
| I | 15 | 1 | 2 | 1 | 0 | 1 | W | 8 | 43 | 37 | 0 | 0 | 9 |
| K | 0 | 0 | 0 | 0 | 0 | 1 | Y | 2 | 36 | 27 | 0 | 0 | 10 |
| L | 6 | 1 | 0 | 5 | 0 | 0 | Total Aligned | 100 | 47 | 53 | 122 | 17 | 50 |
| M | 19 | 2 | 0 | 1 | 0 | 0 | | | | | | | |

1.2288 Å, a PHD finger with an RMSD value of 1.1067 Å, two PWWP domains, with the lowest RMSD value being 1.2739 Å, and a ChromoShadow domain with an RMSD of 1.3865 Å. Although we did not include Tudor domains in our PDB searches, a Tudor domain did show up in the top ten superpositions, with an RMSD of 1.4128 Å. This may be explained by discrepancies between the structure title (PDB searches) and the automatic protein family assignment (pfam family) that we already mentioned. Like for the Tudor domain, we used the superpositions to check the amino acid conservation of HC residues of ARFs that were identified by structural alignments with Tudor domains in the non-Tudor domains (F298, F308, W335, E337, F342, R343) (Table 5.3).

Apart from the RMSD and residue conservation in the top ten lowest RMSDs, the conservation of the selected HC amino acids was also analysed for 140 structures which had a RMSD values lower than 2.0 Å. The most conserved positions were F308 and W335, where the most abundant residues were aromatic (97/136 and 113/131, respectively) (Table 5.4). This is in line with the expected conservation of the hydrophobic cage. However, F298 in AtARF1 seems to be more variable, mostly alternating between Asp, Met, Ile and His in the aligned structures. E337 is mostly occupied by glycine (70/133), followed by an aspartic acid as the second most frequent amino acid (33/133), similar to the glutamic acid found in AtARF1. The prevalence of Gly suggests that the incorporation of an amino acid with a small or absent sidechain may have a specific functionality, for example a spatial tolerance to big ligands. The negatively charged or bulkier sidechains on the other hand suggest a change of the specificity for smaller or basic substrates. F342 was the least structurally conserved of the HC residues, as only 31 out of 140 structures had a residue in a similar position as this residue, indicating that this region is structurally highly variable among the analysed structures. It also suggests that the position occupied by F342, in the loop that is blocking the AD, is in a conformation quite exclusive of ARF-ADs. Surprisingly, the AtARF1 R343 position is mostly occupied in the analysed RF members by aspartate, glutamate, glycine, threonine and tryptophan, which suggest that the blocking mechanism by an amino acid with a long basic sidechain located in the β 3- β 4 loop is mostly exclusive to ARF-ADs.

5.4 Discussion

In this chapter we provide bioinformatic support for the hypothesis that the ARF Ancillary Domain harbours a functional Royal Family domain. We show that the last 80 amino acids of all ARF-DBD structures are structurally similar to the RF family domains discussed in chapter 4. Furthermore, HC residues that are critical for PTM recognition in Royal Family members are preserved in ARF-ADs, but variations in surrounding amino acids suggest differentiation in the recognition of PTM patterns between ARFs AD, for example different methylation states and/or peptide sequences. In addition,

we found that the characteristic Royal Family Hydrophobic Cage responsible of the PTM reading function is actually present in ARF-ADs but that this cage is occluded by a basic residue of the AD. The region that contains the blocking residue in ARF-ADs forms a random coil in ARFs but it is structured as a β -strand in other RF members. It should be noted that the C α atoms of residues in this part of the structure are in proximity in ARF and RF proteins, and it is possible that upon ligand binding, this random coiled loop in ARFs could change conformation to form a β strand that is found in other RF members. The interaction of R343 with the HC would function as a regulatory mechanism of the HC entrance, allowing AD activation only after a conformational change of the loop and a concomitant flip of R343. We have shown that occluded cages also occur in RF members. These are classified as inactive in some cases, whereas in other cases, a movement of this blocking residue has been also demonstrated to occur upon ligand binding. Interestingly, the deletion of just the Ancillary Domain disrupts AtARF5 function *in vivo* [165], which suggests that this subdomain of the DBD, whose function is still unknown [27], is critical for ARF function.

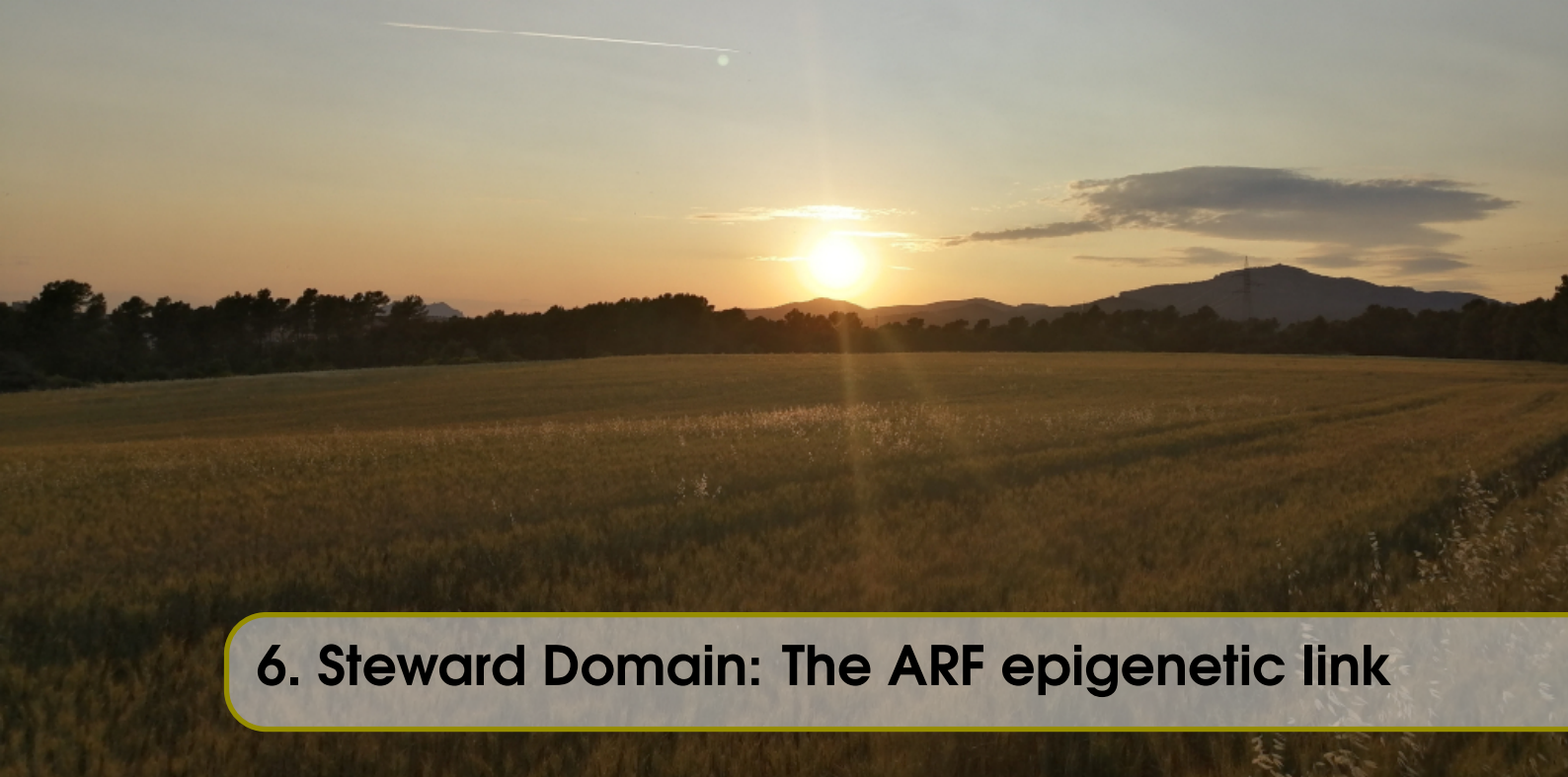
According to our results, the overall structure of ARF Ancillary Domains certainly resembles other Royal Family members but we cannot determine to which family they belong as Ancillary Domains lack structural characteristics of different subfamilies, as well as share structural elements of different subfamilies. The ARF-AD does not form a barrel structure, which is a signature of Tudor, MBT, Chromo Barrel and PWWP domains [99]. In contrast, ARF-ADs possesses a 3_{10} helix at the C-terminus, which is a signature of Chromo- and Tudor-like domains. In addition, the side of the β sheet where the HC is located in the AD resembles that found in chromodomains, but in the former, the β sheet is formed by five strands, whereas in the latter, the β sheet is formed by three strands. Furthermore, no sequence conservation exists between ARF-ADs and any of the RF members, and BLAST searches using AtARF1-AD sequence as query only retrieved ARF sequences. This indicates that ARF-ADs can be classified as a new family inside the Royal Family, which probably diverged from a common ancestor of the chromo-, Tudor, MBT, chromo barrel, and PWWP domains. This divergence is specific of plants, as we were unable to find similar sequences outside of the plant kingdom.

We have also shown that the DNA Binding Domain of ARFs probably carries a second RF-like region within its Dimerization Domain. We propose that this domain does not have PTM recognition function because its hydrophobic cage is stabilizing the $\alpha 6$ helix of ARF-DBD, which is fundamental for ARF-DBD dimerization [39]. Given the fact that two putative Royal Family domains exist in AD and DD, the configuration is very similar to what is observed in extended and tandem Tudor domains, or in the tandem Chromo Barrel domain proposed in the previous chapter. As we discussed, Chromo Shadow domains always occur in combination with an active chromo domain, and Chromo Shadow domains have an intrinsic ability to self-dimerize [94, 112]. This is similar to what occurs in ARF-DBDs, where the domain found in the DD would act as a dimerization spot, with no histone PTM binding ability, similarly to the Chromo Shadow domains. The Ancillary Domain would act as the histone PTM binding site and could represent a fully functional RF member. A similar situation is observed in extended Tudor domains, where the ARF RF-like domain in the DD would be equivalent to the extended Tudor SN-like domain, and the ARF AD equivalent to the active Tudor domain. This supports the hypothesis of an ancestral predecessor of both ARF-ADs and Royal Family members with a combined active and inactive RF domain. In several of these combined RF domains analysed in this thesis, the PTM binding requires both domains [100, 122], so this requirement should be tested in ARFs. Interestingly, the B3 domain seems to be an insertion in an ancestral tandem Royal family domain, which evolved jointly with Ancillary Domain.

As we will cover in the next chapter, AtARF1-DBD is able to recognize posttranslational

modifications in peptides derived from histone tail sequences. For this reason, we believe that this type of domains should be included in a separate family within the Royal Family, in addition to the existing families. We propose the name “**Steward domain**” for the domains that are similar in structure and sequence to ARF DD-AD domains. We chose this name to continue the naming scheme used for domains of the RF, which is based on English monarch dynasties, as exemplified by the plant Agenet and Tudor domains. The House of Stewart ruled after the Tudor dynasty, and we thought it appropriate to substitute the last “t” by a “d”, because of the relation between the word steward and the original name for the Ancillary Domain on the one hand, and the assisting, regulatory function of R343.

The possibility that the ARF-DBDs harbour a functional PTM reading domain in their DBDs suggests that gene selection is based on their ability to bind histone tails and thereby establish a direct interaction with nucleosomes. This hypothesis connects epigenetics and auxin sensing and has the potential to explain how the highly conserved ARF-B3 are able to regulate a wide variety of responses in the presence of auxin. In this scenario, ARF will mediate the crosstalk between changes in epigenetic marks and AuxRE availability, interconnecting direct DNA recognition and PTM selection, which has the potential of fine-tuning expression of auxin-dependent genes. If true, this not only entails a completely new aspect of auxin biology and gene regulation, but also makes the AD a target for developing specific interactors, which could function as growth regulators or herbicides. Several initiatives for developing specific chemicals against other Royal Family members have emerged to regulate cellular growth, such as in cancer therapies in humans [117, 166, 167, 168]. A similar approach could be applied to ARFs, where molecules are designed to directly alter PTM recognition of ARF-ADs or target the enzymes that add or remove certain PTMs to alter ARF binding, resulting in a final tuning of ARF-regulated gene expression.



6. Steward Domain: The ARF epigenetic link

6.1 Abstract

The structural analysis of the ARF-DBD identified a putative functional Royal Family-like module within the ARF-DBD. The presence of an histone PTM reader module in ARF-DBD would allow ARFs to bind to certain histone modifications, thus modulating the binding of ARFs to the target sites. Under this hypothesis, ARF proteins would integrate both the epigenetic information present in chromatin as well as the auxin input to a combined response on gene expression, thus affecting the cellular fate. While this hypothesis is exciting and would help explain the complexity of ARF signalling despite the small number of components on the auxin transcriptional response pathway and the sequence conservation on DNA-interacting residues, no experimental evidence of a functional ARF histone PTM reader module has been reported. In this chapter we report *in vitro* experiments that are consistent with a functional RF-like module in ARF-DBD, and that this module is subject to regulation. All the information gathered in this chapter lays the foundations of new *in planta* experiments to further test the importance of ARF Steward domain on auxin signalling.

6.2 Introduction

Structural and bioinformatical alignments using the Ancillary Domain, shown in previous chapters, suggest that ARF-DBD contain a domain that may be classified within the Royal Family of histone posttranslational modification readers, which we called the Steward domain.

The study of histone PTM readers usually requires prior information to determine the peptide sequence and modification for confirmation of the interaction with a certain peptide. This information can be gathered from several protein-protein interaction techniques that can be based on known partners of the domain [37], from homology [169], or from *in vitro* tests incubating cellular

extracts with the protein of study as bait, in an attempt to catch interactions in assays known as “pull-down” [170]. While these experiments give valuable information on the interactors of the protein, they may miss hits due to low affinity of the interaction, the requirement of activation or modification of the interacting or the study protein, or the requirement of gene expression of the interactor. As we recently determined that ARFs contain the putative PTM Steward domain, and there is no information on the function of this domain or possible interactors, trying all the possible combinations of peptide sequences and modifications may overcome the lack of ARF Steward domain information.

Given the vast number of histone types, variants, regions of modifications and type of modifications, trying each of the modifications for ARF-DBD binding, one by one, is unfeasible. In recent years, the proliferation of “omic” sciences (genomics, proteomics, interactomics, metabolomics, ...) led to the development of several techniques with high throughput capabilities [171, 172, 173, 174, 175, 176]. On the one hand, techniques with the ability to probe for binding partners of a certain protein (interactomics) were developed, using the protein of interest as bait for picking interactors from complex samples. One of these techniques, **pull-down assay**, is an easy to implement technique which can detect protein-protein interactions. The problem, though, is that weak, transient interactions and interactions that only occur after an activation event will be missed, so many other strategies are used to improve the detection of the interactors [170, 177].

A huge development was the **array assay**. Thanks to the many different analytes that an array can hold, arrays can be used in genomics with DNA libraries; in proteomics, with peptide libraries; in ligand and drug discovery, where collections of small molecules are tested for binding with the protein of study, to mention a few of many other possible applications. The development of array-based techniques was especially useful in genomics, as the four nucleotides that compose the genetic code allow for the random synthesis of all the possible sequences for a given length [62, 178]. For example, all the possible nucleotide sequences with a length of ten nucleotides amount to roughly 10^6 , a number suitable for high throughput techniques. In contrast, as proteins are composed of twenty amino acids, the number of possible sequences is greatly surpassed for a peptide of just five amino acids in length. Given that protein interactions usually require longer peptides and may involve posttranslational modifications, the use of random sequences to probe the possible peptides that interact with a protein is normally not a viable strategy. Thus, a different approach is required that avoids the use of random sequence libraries. One of the strategies is the use of peptide libraries associated with a certain area of study, which can be thought of as a contextual library. Contextual libraries contain a combination of peptides related to a certain disease, cellular status or protein function. Since bioinformatic results suggested that the ARF Steward domain may bind posttranslationally-modified histones, we aimed to probe whether ARF-DBDs were able to bind other proteins using pull-down assays, or if the ARF-DBDs were able to bind histone PTM using peptide libraries of histone peptides with different posttranslational modifications. However, due to the complexity of the molecular and environmental circumstances, no single omics analysis can independently explain the intricacies of fundamental physiology [174], so we required parallel experiments to confirm the findings.

6.3 Results

6.3.1 The Ancillary Domain alone is not sufficient for PTM binding

In order to determine the binding partners of ARF-AD by pull-down assays, significant quantities of pure His6-tagged protein are required. For this purpose, we cloned, expressed and purified the Ancillary Domain of MpARF2, residues 318-397. Although it was possible to obtain sufficient His6-tagged protein for the experiments, the yield of the purification as well as the stability of the protein were significantly lower than in the whole ARF-DBDs. The obtained protein was used for pull-down assays, where extracts of *Marchantia polymorpha* plants were incubated with MpARF2-AD bound to His-Trap resin, and after several washes, the His-Trap bound fraction was analysed by SDS-PAGE. No significant interactions were detected by Coomassie staining (Figure 6.1), so a different approach was followed. In this second approach, we enriched *Marchantia polymorpha* plant extracts by pull-down, to then denature and separate these enriched extracts by SDS-PAGE. The resulting separation was transferred to a PVDF membrane and blocked like in a Western Blot, following an Overlay Blot protocol [177]. The membrane was later incubated with recombinant N-ter His6 tagged MpARF2-AD and after successive washes, the membrane was revealed for the presence of the His tag. The result shows that a band of 10-15KDa and other several less intense bands appear in test lane that are not present in the control lane (Figure 6.2). The identity of the main 10-15KDa band is still under investigation.

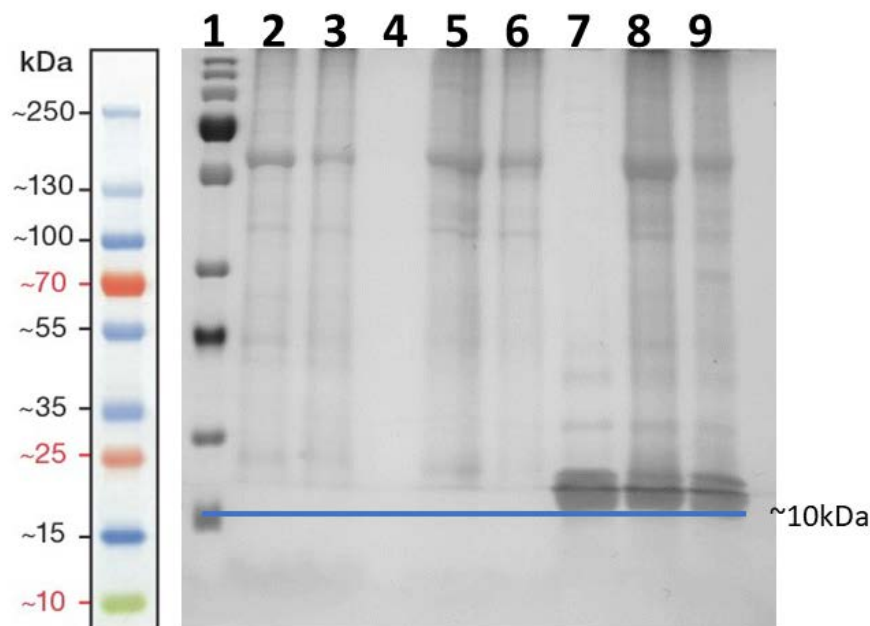


Figure 6.1: Pull-down assays with purified MpARF2-AD. 1: Marker, 2: fresh *M. polymorpha* extract, 3: 24h old *M. polymorpha* extract (not frozen), 4: Beads, no extract, 5: Beads incubated with the extract loaded in lane 2, 6: Beads incubated with the extract loaded in lane 3, 7: Beads incubated with MpARF2-AD, 8: Beads incubated with MpARF2-AD and the extract of lane 2, 9: Beads incubated with MpARF2-AD and the extract of lane 3.

Given that the results of these techniques were not conclusive, other protein-protein interaction assays were performed to further test the ARF-AD interaction profile. As we commented earlier in chapters 4 and 5, most Royal Family domains that are found in tandem require both domains of

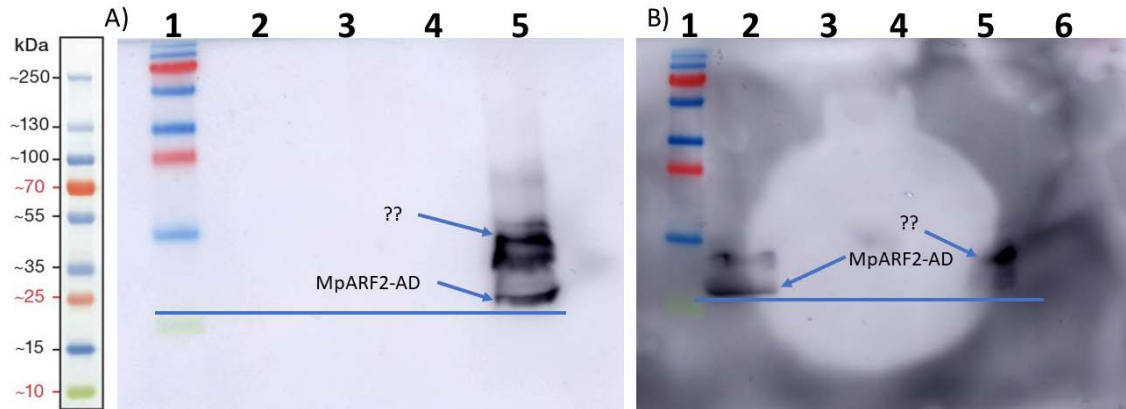


Figure 6.2: Overlay blot of *Marchantia* extracts using enriched samples by pull-down followed by detection with MpARF2-AD 6His + α 6his + primary Ab. A) 1: Marker, 2: Supernatant of MpARF2-AD after bead incubation, 3: Supernatant of the extract incubated with the beads, 4: Supernatant of the extract incubated with beads loaded with MpARF2-AD. 5: boiled beads incubated with MpARF2-AD and the extracts. B) Same as A), but with MpARF2-AD control. 1: Marker; 2: MpARF2-AD, Supernatant of beads incubated with MpARF2-AD; 3: Beads control; 4: Supernatant of the extract incubated with beads; 5: Boiled beads incubated with MpARF2-AD and extract; 6: Supernatant after bead incubation with MpARF2-AD and the extract. An increment in a \approx 15kDa band compared with control is observed.

the tandem for a proper function. For this reason, we decided to test the whole AtARF1-DBD to consider the bioinformatical suggested possibility that the Ancillary Domain and the secondary Royal Family domain located in the Dimerization Domain form the aforementioned Steward domain.

6.3.2 AtARF1-DBD as an effective histone PTM reader

To determine the binding preferences of ARF-ADs, a library of peptides was used (Histone Code Peptide Microarrays, JPT Peptide Technologies GmbH, Product code: His_MA_01), which contained peptides corresponding to the tails of the four core histone subunits plus the H1 linker histone, and which also contained several modifications on Arginine, Lysine, Serine, Threonine and Tyrosine residues, including acylations, methylations, phosphorylations, among others, at different positions and combinations [179]. The large number of different peptides (more than 3800) allowed us to thoroughly test the whole landscape of the binding preferences of AtARF1-DBD using a microarray-based assay. It is worth noting that the histone sequences used by the manufacturer correspond to human sequences. Although histones are highly conserved during evolution, we checked the main sequences and the differences found between array and *Arabidopsis thaliana* sequences are considered and discussed in the interpretation of the result.

The top ten strongest binding peptides to AtARF1-DBD were H2A 71:90 R81me1, H2B 81:100 pS87, H1 71:90 K84prop, H3 63:82 K79me3, H2B 81:100 R92me2s, H3 11:30 K18but, H2A 25:44 R39me1, H2A 28:47 K37me2, H2A 11:30 R20me2s and H2A 69:88 K74suc. These peptides reveal a tendency of AtARF1-DBD to interact with modified arginines and lysines. Thus, mono and symmetric dimethylated arginines are overrepresented (5/10) as are acylated lysines (which include propionylation, trimethylation, butyrylation and succinylation) (4/10). For all these peptides, we contrasted their binding affinities with that of all the PTM present in the array for that

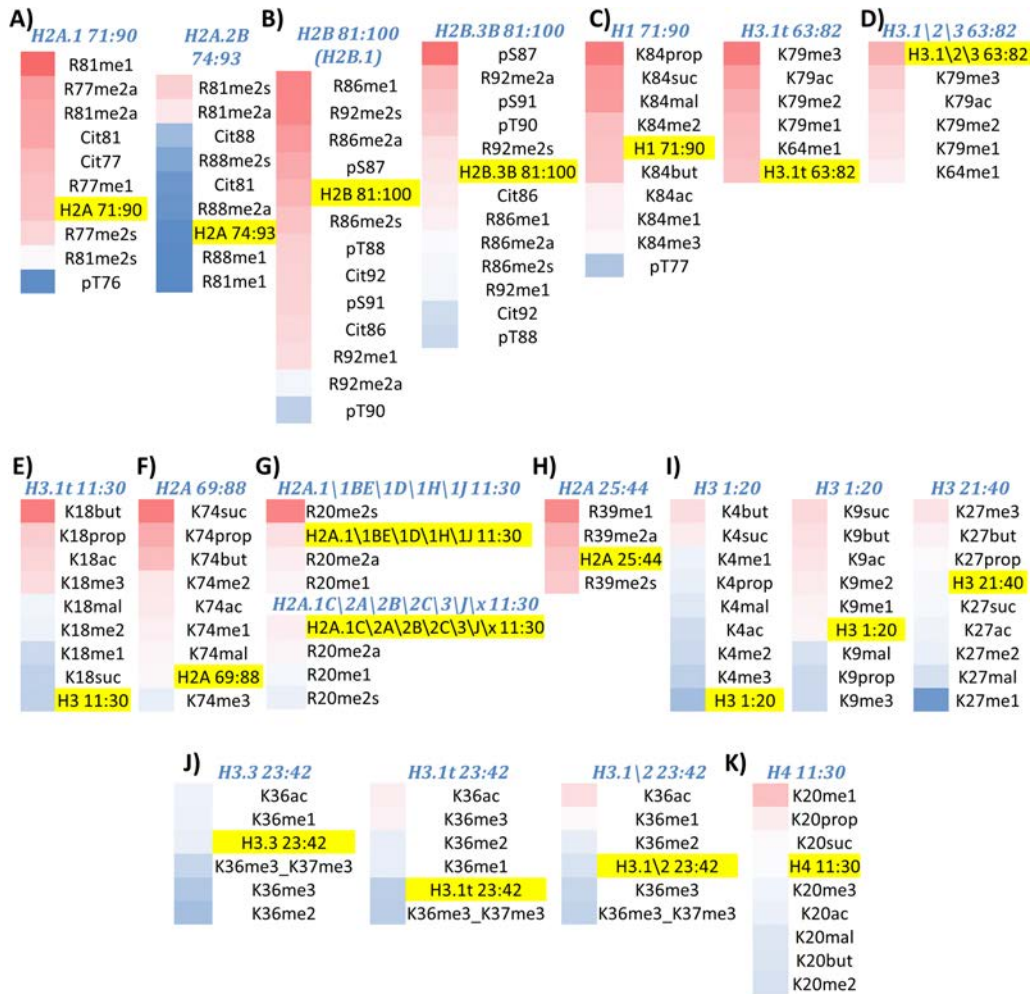


Figure 6.3: AtARF1-DBD interactions with Histone Code Peptide Microarray for the ten strongest interactions and for peptides known to bind the RF. The colour scale is relative to the maximum (Red) and the minimum (Blue) interaction affinity in the whole array. The colour bands on the vertical bars represent the binding intensity of all of the modifications for the peptide with highest affinity, shown above the bar. Peptide labels highlighted in yellow correspond to unmodified peptides.

particular peptide sequences. In addition we specifically checked the affinities of AtARF1-DBD for peptides that are known to bind to Royal Family domains, i.e. TGGVKKPHR (H3 32:40 K36), ARTKQTARKS (H3 1:10 K4), KRHRKVLDRN (H4 16:25 K20) and IAQDFKTDLRF (H3 74:84 K79), and all the PTMs thereof included in the array (Figure 6.3).

The strongest interaction found was with the peptide H2A.1 71:90 R81me1. Comparison of the results obtained with peptides containing the H2A.1 71:90 sequence (RDNKKTRI-IPRHLQLAIRND) and peptides containing the H2A.2B 74:93 sequence (KKTRIIPRHLQLAVRN DEEL), including all available PTMs in the array (Figure 6.3A), reveal that AtARF1-DBD in general has preference for the peptides containing I87 (H2A.1 71:90). Thus, AtARF1-DBD has higher affinity for the unmodified H2A.1 71:90 peptide than for the unmodified H2A.2B 74:93 peptide. Interestingly, AtARF1-DBD shows diverse histone variant selectivity for H2A.1 and H2A.2B variant, where a preference was seen for R81me1 in the former and R81me2 in the latter, although the interaction with H2A.1-based peptides is always stronger. We scanned *Arabidopsis*

thaliana protein sequence databases in search for homologs of the human sequences H2A.1 71:90 and H2A.2B 74:93 of the array. While we found that the H2A.1 71:90 sequence (Isoleucine in position 87) was totally conserved, we were unable to find sequences with a Valine in position 87. Taking this in consideration, the lower affinity of AtARF1-DBD towards the H2A.2B 74:93 sequence makes sense because *Arabidopsis thaliana* does not encode for histones with the I87V substitution. In any case, the high affinity for R81 symmetric and asymmetric dimethylation in both H2A/2B and H2A/1 sequences shows that the protein is specific for this modification.

Histone H2B.3B 81:100 pS87 (AHYNKR-pS-TITSREVQTAVRL) showed the second most intense interaction of the whole array. This peptide is similar to histones H2B.1\1B\1D\1H\1J\1K\1L\1M\1N\1O\2E\2F, which are represented by peptide H2B 81:100 R86me1 in position 7 (AHYNKRMe1-STITSREIQTAVRL) and peptide H2B 81:100 R92me2s (AHYNKRSTITS-RMe2s-EIQTAVRL) in position 9 of the ten strongest interactions. The sequence of H2B.3B 81:100 and H2B.1 only differ in the occurrence of an Isoleucine (H2B.1) or Valine (H2B.3B 81:100) at position 94. This change of a single amino acid leads to differences in the recognition of the PTM pattern (Figure 6.3B). For peptides carrying V94 (H2B.3B), AtARF1-DBD was preferentially interacting with phosphorylated serine 87 and, to a lower extent, with asymmetric dimethylated arginine 92. The preferred interactions of the protein with pS91 and pT90 PTM variants of this peptide indicate that phosphorylation help establish interactions with AtARF1-DBD. Peptides with the Isoleucine at position 94 have different interaction strengths, even though the sidechains of valine and isoleucine are similar in length and properties. For these peptides, pS87 still increases the binding, but now the AtARF1-DBD preferentially interacts with monomethylated and asymmetrically dimethylated R86, and symmetrically dimethylated R92. The peptide sequence containing I94 is more relevant physiologically, as the sequence found in *Arabidopsis thaliana* (**ARYNKKPTITSREIQTAVRL**) contain this Isoleucine. The bold residues represent the residues in *Arabidopsis* that are different with respect to the human H2B.3B 81:100 sequence which can affect the affinity for the protein, so further testing is required to confirm whether AtARF1-DBD has affinity for the H2B histone from *Arabidopsis thaliana* with monomethylated and asymmetrically dimethylated R86 or symmetrically dimethylated R92.

The third most intense interaction corresponds to the peptide H1 71:90 K84prop (IKRLVTTGVLKQT-Kprop-GVGASG). Histone H1 has a role in chromatin compaction but is the less studied histone. Thus, the functional implications of the interaction of AtARF1 with H1 cannot be established. The affinity of dimethylation of K84 of this peptide slightly increases the affinity, but the most prominent effect is produced by propionylation of this residue. Succinylation and malonylation also increase affinity albeit to a lesser extent, while butyrylation and acylation reduces affinity. Malonylation and succinylation, which introduce an acidic group, also increase affinity [180, 181]. Finally, phosphorylation of T77 reduces the interaction with the peptide (Figure 6.3C). The possible implications of these interactions in *Arabidopsis thaliana* are not clear, as we were unable to find an *Arabidopsis thaliana* histone with similar sequence to human H1.

The fourth highest affinity was observed for the H3.1t 63:82 peptide with a trimethylated K79 (RKLPFQRLMREIAQDF-KMe3-TDL). The H3 63:82 region harbours one of the typical Royal Family interacting peptides presented before (H3 74:84 K79, IAQDFKTDLRF), with the modification K79. The interaction strength does not change much for the tested modifications, although this histone variant H3.1t shows increased affinity with increasing methylation. In contrast, in peptide variants based on the H3.1/2/3 63:82 sequence (RKLPFQRLVREIAQDFKTDL), which contain a M71V substitution, K79 methylation reduces the affinity of the peptide (Figure 6.3D). Considering that the H3.1/2/3 sequence is completely conserved in *Arabidopsis thaliana* and that we were unable to find histone sequences containing residue M79 in *Arabidopsis thaliana*,

AtARF1-DBD may interact with the H3.1/2/3-like sequences *in vivo*.

The fifth most intense spot in the array corresponds to H3 11:30 K18but (TGGKAPR-KBut-QLATKVARKSAP) (Figure 6.3E). AtARF1-DBD displayed similarly strong interactions with K18 trimethylation and several acylations. Other modifications result in weaker interactions, although all the assayed modifications result in higher affinity compared to the unmodified peptide.

The H2A 69:88 K74 succinylation (AGNAS-KSuc-DLKVKRITPRHLQL) is the sixth most intense interaction on the array. Thus, similar to peptides H3 11:30 and H1 71:90, lysine acylation resulted in a stronger interaction compared to the unmodified peptide. Dimethylation and monomethylation of K74 increased the binding affinity compared with the unmodified peptide, whereas trimethylation slightly reduced it (Figure 6.3F). This sequence was found completely conserved in *Arabidopsis thaliana* H2A.

The eighth most intense interaction in the array involved the H2A 11:30 peptide with asymmetric dimethylation of R20 (RAKAKTRSS-RMe2s-AGLQFPVGRV). This peptide is derived from the histone variant H2A.1\1BE\1D\1H\1J with sequence (RAKAKTRSSRAGLQFPVGRV). The modification of R20 increases the affinity substantially compared with the unmodified peptide and other modifications. For the very similar H2A.1C\2A\2B\2C\3\J\X variant (RAKAKSRSSRAGLQFPVGRV), the affinities were low for all PTM variants. Thus, the T16S substitution abolishes the interaction affinity of AtARF1 (Figure 6.3G). The most similar histone sequence present in *Arabidopsis thaliana* (KAKTKGKSRSSRAGLQFPVGRI) contains the S16 residue, so it is likely that this interaction is not produced *in vivo*.

The tenth most intense interaction on the array corresponds to the H2A 25:44 peptide containing the monomethylation of R39 (LQFPVGRIHRHLKS-RMe1-TTSHG). A slight increase in interaction affinity is observed compared to the unmodified peptide. The other modifications of this peptide do not significantly alter the binding compared to the unmodified peptide (Figure 6.3H). The sequence found in *Arabidopsis thaliana* differed significantly in the region involving R39 (LQFPVGRVHRLKTRSTAHG), so it is not clear whether an interaction is possible *in vivo*.

As we commented above, peptides H2B 81:100 R86me1 and H2B 81:100 R92me2s were ranked in 7th and 9th position, respectively. Thus, three of the ten strongest interactions of AtARF1-DBD occur with modifications in the H2B 81:100 peptide (AHYNKRSTITSREIQTAVRL) containing methylated arginines. Further investigation is required to confirm the high preference of AtARF1 for this peptide by other techniques, as the results may indicate H2B as a binding partner of AtARF1 with variable affinity depending on the modification.

We have also specifically checked the affinity of AtARF1 with sequences that are recognized by RF family domains [115, 117, 121, 140, 182, 183, 184, 185, 186], which include the H3 and H4 sequences. In Figure 6.3I we show the interaction intensity of AtARF1-DBD with modifications localized in H3 1:40 region H3K4, H3K36 and H3K79 and in the H4 11:30 K20. Most of these are weak AtARF1-DBD interactors, but the interactions observed are strongest for lysine acylation, as was also observed in the case of H2A 69:88, H3 11:30 and H1 71:90 (see above). In H3 1:20, peptide, butyrylation and succinylation of lysine 4 or 9 resulted in the strongest interactions, whereas H3K36 modifications resulted in poor interaction with AtARF1-DBD, with only K36 acetylation slightly increasing the affinity (Figure 6.3J). As we already commented, the peptide H3 63:82, which is another known Royal Family interactor, resulted in the fourth most intense interaction for the modification in which K79 was trimethylated (Figure 6.3D). Thus, the AtARF1-DBD can have affinity for peptides known to bind to RF. Finally, the peptide H4 K20, which is the most

studied Royal Family substrate, shows affinity to AtARF1-DBD for variants in which K20 is monomethylated, whereas dimethylation and trimethylation reduces affinity in comparison with unmodified peptide (Figure 6.3K). The sequences of these four H3 and H4 derived peptides are completely conserved in *Arabidopsis thaliana* histones.

6.3.3 AtARF1-DBD binds RF substrates with low micromolar affinity

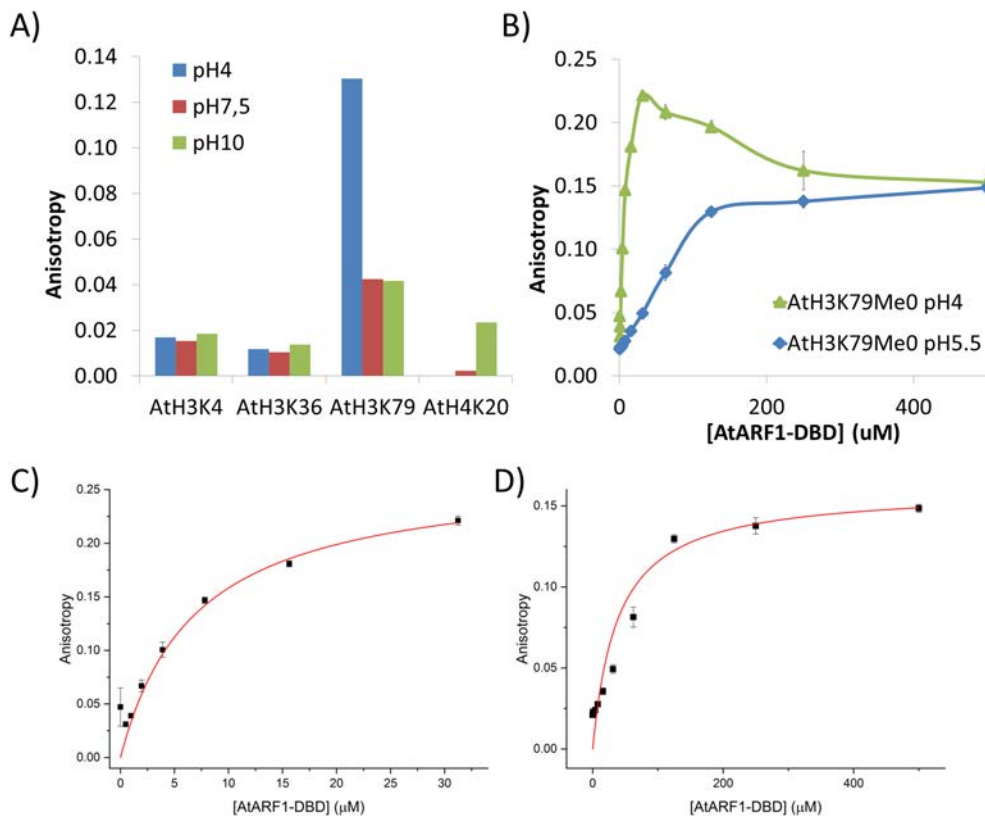


Figure 6.4: Quantitative analysis of the interaction between AtARF1-DBD and AtH3 74:84 peptide using FA. A) Fixed concentrations of 6-Fam labelled peptides (50 nM) and AtARF1-DBD (500 μM) were used to screen peptide interaction at different pHs. B) Comparison of the anisotropy produced in 6-FAM AtH3 74:84 peptide at different AtARF1-DBD concentrations ranging from 0.5 to 500 μM. C) and D) show the fit to the Michaelis-Menten equation of the AtARF1-DBD concentration points in the hyperbolic region of the curve at pH 4 with a calculated $K_D(pH4) = 6.91 \pm 0.83 \mu\text{M}$ (B, 0.5-31.25 μM) and in all the concentration region of the curve at pH 5.5 with a calculated $K_D(pH5.5) = 38.73 \pm 9.55 \mu\text{M}$ (C, 0.5-500 μM).

The array results suggested several candidates as possible interactors of the AtARF1-DBD. To confirm these results *in vitro*, we used fluorescence anisotropy to measure the affinity for AtARF1-DBD of selected peptides. Given the high number of positive hits, we prioritized the sequences where other Royal Family members showed interactions in X-Ray structures. The assayed canonical histone peptide sequences were unmodified H3 1:10 (ARTKQTARKS), H3 32:40 (TGGVKKPHR), H3 74:84 (IAQDFKTDLRF) and H4 16:28 (GGAKRHRKVLDRDN).

Because the structural analyses showed that the putative HC of the Ancillary Domain is blocked by R343, we decided to test whether the pH has an effect on the binding affinity of the peptides.

Thus, we determined the binding potency of these peptides at different pHs at saturating AtARF1-DBD concentrations. Results at pH7.5 show that peptide H3 74:84 is the only that interacts to a higher extent compared to the other peptides, where the other peptides showed similar anisotropy values at all pHs tested were obtained (Figure 6.4A). The interaction of H3 74:84 is greatly enhanced at pH4, with a slight interaction higher interaction compared to the other peptides at pH 7.5 and 10. This indicates that the pH affects the interaction, and although a pH of 4 does not occur under physiological conditions, it suggests that regulation of the interaction may occur through local proton increment, ligand concentration or ARF Steward domain activation by PTMs.

Saturation curves at pH 4 with the AtH3 74:84 peptide show a Michaelis-Menten behaviour up to an AtARF1-DBD concentration of 31.25 μM (Figure 6.4B). After reaching this point, the anisotropy started to drop until a plateau is reached between 125-500 μM . Our explanation to this behaviour is that the extreme conditions produced by pH4 induce aggregation and inactivation of AtARF1 at high concentrations, as we observed white precipitates at the highest concentrations after performing the measurements. This white precipitate was not observed at pH 5.5, reinforcing the explanation that the precipitation could be the reason of the inactivation observed at pH4. Despite the precipitation, the first data points in the 0-31.25 μM range provide interesting information. First, the maximum anisotropy achieved at pH4 is almost 50% higher than at pH5.5, indicating that the maximum interaction is not achieved at pH5.5. Also, the affinity at pH4 is much higher than at pH5.5, increasing from $38.73 \pm 9.55 \mu\text{M}$ at pH5.5 to $6.91 \pm 0.83 \mu\text{M}$ at pH 4 (Figure 6.4C and D). These K_D s are among the lowest of the reported K_D s of Royal Family domains [99, 116, 123, 125, 140, 159, 167, 187], reinforcing the notion of a histone PTM reader domain is located in the Steward domain. The pH dependence also suggest that the binding is subject to regulation, and that the inactivation follows a mixed inhibition mechanism, as the maximum binding decrease and dissociation constant increase with increasing pH [188]. These results show that at least for the *Arabidopsis thaliana* H3 peptide, no strong interaction can be expected at physiological pHs, unless an activation of the domain is produced. It should be pointed out that this peptide is unmodified, and that modifications that increase significantly the affinity of this peptide for AtARF1-DBD were suggested by microarray assays. Future research will be conducted with the peptides found on microarray assays.

The quantitative results obtained by fluorescence anisotropy are in good agreement with the results obtained from the microarray experiments, even though the microarray experiments were performed at physiological pHs. Fluorescence anisotropy shows that of the four tested peptides, derived from H3 and H4 sequences, only the H3 74:84 peptide displayed a high interaction affinity, in line with the observed affinity of the array peptide H3 63:82. In the case of H4 16:25, we did not see interaction in FA at any pH, whereas a strong interaction is observed for the monomethylated variant of K20 of the H4 11:30 peptide in the microarray experiment, suggesting that this interaction is dependent of the monomethylation for the binding. Future binding assays will include this modification to corroborate this finding. Finally, no affinity was detected for the H3 32:40 and H3 1:10 peptides in fluorescence anisotropy, which agrees with the microarray results obtained for the H3 23:42 and H3 1:20 peptides. Since the affinities of the microarray experiments are in good agreement with the affinities measured using FA for the peptides recognized by RF family domains, the interaction affinity of the most potent peptides found in array assays should also be quantified. If confirmed, this would represent new specificities inside the Royal Family of domains.

6.4 Discussion

Following up on bioinformatic analyses of the Ancillary Domain, we have tested the binding affinities of AtARF1-DBD to modified histone peptides in this chapter. We provide the first experimental evidence of a histone binding function in ARF-DBD.

The first attempts to find interaction partners of MpARF2-AD resulted in inconclusive results. For these trials, pull down assay and overlay blot assays using purified MpARF2-AD were employed. As has been suggested in previous chapters, the ARF-DBDs may harbour an additional RF-like in the DD, which probably is required by ARF-ADs to function. Thus, we focused our efforts on the detection of possible ARF-DBD and ARF-AD histone PTM binding. However, the stability and purification yield of MpARF2-AD was very low, which complicated the experiments. For this reason, we opted to use the whole AtARF1-DBD to test the histone PTM binding ability of ARF-DBDs.

By assessing binding of AtARF1-DBD to an array of modified histone peptides, we have mapped the specificity of AtARF1-DBD for different types of histone peptides with an array of posttranslational modifications. AtARF1-DBD preferentially recognized arginine methylation in H2A R39, H2A.1 R20/R77/R81 and H2B R86/R92. Lysine modifications that resulted in binding of AtARF1 were primarily acylations, in contrast with what is known for other Royal Family members, none of which has been reported as a lysine acylation reader. These results are very interesting because they clearly point to a specificity of AtARF1-DBD for H2A/B arginine methylation, and that AtARF1-DBD would also be able to discern between lysine acylations. Furthermore, AtARF1-DBD is able to discriminate between histone variants of H2A, H2B and H3, where single amino acid substitutions in the surroundings of the modified residue produced a shift in the recognition of the PTM patterns. This shows that posttranslational modifications in different histone variants completely alter the affinity for AtARF1-DBD. It is known that switching from canonical histones to other histone variants alters nucleosome stability and helps to create functionally distinct chromatin domains [96, 189, 190]. It has been proposed that different histone variants regulate distinct cellular processes by differential recruitment of binding partners, such as DNA repair and transcriptional activity [189]. Thus, the histone composition of nucleosomes determines their properties and their interactions with chromatin remodellers and modifiers. For example, H2A variants affect gene expression: H2A.Z and H2A.B are implicated in transcription initiation, whereas macroH2A in animals [191] and H2A.W in plants [192] seem to be associated with nucleosome immobility and transcriptional silencing [193]. The molecular mechanisms that govern nucleosome modification are known, but the mechanism of transcription regulation by histone variants incorporated in nucleosomes is the subject of considerable ongoing study [81]. Here we report the effect of point substitutions found in histone variants of H2A, H2B and H3, which are located in regions near the assayed PTMs on their binding affinity towards AtARF1-DBD. Most of the known histone variants are homologs of H2A and H3 histones [192], and the amino acid changes present in histone variants near the assayed PTMs could alter recognition and, as we have shown, reconfigure the substrate specificity of AtARF1-DBD. More in depth studies will be required to clarify if histone variants with substitutions in the vicinity of modified residues also modulate the binding affinity of other histone-binding proteins.

The array analysis of the affinity of typical Royal Family substrates for AtARF1-DBD showed that their affinity was much lower than that of the H2 substrates, with the exception of peptide H3.1t 63:82 K79me3. Generally, H3K4, H3K36 and H3K79 methylations are considered to mark active transcription, whereas H3K9, H3K27 and H4K20 methylations are generally associated with silenced chromatin states [194]. Our array results show poor or no interaction of AtARF1-DBD

with H3K4 and H3K36, but high affinity for H3K79. From these results we cannot infer a direct relationship between the binding affinity and a hypothetical activation/repression activity of the AtARF1 with the canonical RF binders, so further research is required to establish the correlation of AtARF1 gene activation or repression and a given histone modification.

The array analysis also served to assess the kinship of ARF-DBD with Royal Family domains. Generally, Tudor domains recognize methylated lysines and arginines, while chromodomains are exclusive methyllysine readers [195]. The PWWP domains are also methyllysine readers but, unlike the other RFs, are able to bind DNA [125]. MBT domains recognize mono- and dimethylated lysine at a number of different positions on histone H3 and H4 tails [124], and Chromo Barrel domains interact with methylated lysine [106]. In our arrays, AtARF1-DBD preferentially recognized methylated arginine as was the case for Tudor domains, although AtARF1-DBD also recognized histone acylation, which are relatively recently discovered histone modifications [180, 181]. This represents a difference with already known Royal Family domains, which are mainly methylation readers. Domains with histone acylation recognition have been reported before, for example the Yeats Domain [196] or bromodomains [197], but these have no structural similarity with the Steward domain found in ARF-DBDs. Although those domains contain a HC to bind acylated substrates, the barrel-like structure of Yeats domains consists of more β strands than that of the RF domains, and bromodomains have an all- α structure.

It should be noted that the commercial array used in the experiments described above contain peptides with sequences derived from human proteins. Despite the conservation of histones during evolution, several differences between human and *Arabidopsis thaliana* histones were found, also in the peptides that have affinity for AtARF1, as we pointed out above. This reinforces the need to conduct further studies to fully characterize the ARF-Histone binding pattern to establish the relationship between the ARF and its' activating or repressive potential.

We used fluorescence anisotropy to further characterize and quantify the interaction between AtARF1-DBD with several standard Royal Family-recognizing peptides that resulted in an interaction, even if very weak, in the array experiments. These tests showed that only H3 74:84 showed affinity for AtARF1 and that the interaction with this peptide is affected by pH, which is in line with the hypothesis that the arginine covering the HC of AtARF1-AD functions as a regulator. Fluorescence anisotropy results contrast with the array results as an acidic pH was required for detecting interactions in fluorescence anisotropy, while array binding experiments were performed at pH7.5. We can provide three explanations for this divergence. First, the detected interaction in fluorescence anisotropy between H3K79 and AtARF1-DBD was also observed in the microarray experiments at pH7.5, but it was not among the most intense interaction on the array. Similarly, the affinity determined in FA experiments was also weak at H3K79. Second, the peptides used for FA were unmodified, and we have found that AtARF1-DBD preferentially bind methylated or acylated peptides. For this reason, the binding of an unmodified peptide may be less favourable and a putative stabilization of the open conformation of the HC by pH may be required for the binding of the unmodified peptide. Finally, the local peptide concentration achieved in the array may be orders of magnitude higher than that used on fluorescence anisotropy (50 nM), favouring ARF-AD reorganization and interaction, which aids in the detection of an interaction. In the cellular nucleus, a high concentration of proteins may alter the binding properties of surrounding molecules (Macromolecular crowding concept [198, 199, 200, 201]). Thus, fluorescence anisotropy assays support the hypothesis that arginine protonation will reduce its ability to bind to the main chain of E337, P338 and S339 residues, as results show a strong correlation between decreasing pH and increase in the affinity of DBD for the peptide. We recognize that the pH used in these experiments is unlikely to occur in the cellular nucleus, but the pH dependent changes in ARF

binding affinity towards the histone tails could be related to changes in the conformation of the blocking residue, which may represent a putative regulation mechanism that is perhaps dependent on post-translational modifications of residues of the ARF Steward domain. The histone binding function of AtARF1-DBD, although the HC is covered in ARFs, is further supported by the fact that the affinity of AtARF1-DBD to AtH3K79, with a K_D of $6.91 \pm 0.83 \mu\text{M}$, is one of the most potent interactions found for interactions between Royal Family members and histone peptides, which have a typical K_D of $500 \mu\text{M}$ with a few exceptions at $10 \mu\text{M}$ [99, 116, 122, 123, 125, 140, 159, 167, 187].

As we showed in chapter 4, previous studies reported occlusion of complete HCs. The inability of these modules to bind peptides was attributed to these blocking residues [106, 109, 116, 156, 158, 159]. It is possible that these residues are internal regulators equivalent to the AtARF1 arginine 343, and that activation is required to free the RF HCs. There is also a case where a HC tryptophan residue is blocking the entrance in the apo state and is moving on peptide interaction [157]. Despite the existence of HC-blocking residues in RF family members, the structural rearrangements upon peptide binding that are described so far are relatively small compared to the substantial rearrangement that is required for activation of the ARF Steward domain HC. Here we provide indirect evidence for this rearrangement in AtARF1, which could be also important for other Royal family members which have been categorized as non-functional. Tudor domains are sometimes found as tandem Tudor domains where one of the domains recognizes histone tails and the other does not, the latter of which contain complete HCs, but their function is still unknown [100, 116]. Our results open the door to the possibility that secondary, non-functional Tudor domains could be in fact functional under some circumstances, and most importantly, subjected to regulation.

The results described in this chapter support the bioinformatic findings presented in the previous chapter. The function of a new domain is usually inferred from overexpression assays, sequence conservation or by comparison with other similar protein functions (Homology-based prediction) [202]. In the case of ARF Steward domains, sequence conservation did not provide information on the function of this domain, so we inferred the function from the structural similarities with known domains (Structure-based function prediction) [202]. This strategy of inferring function from structure relies on the assumption that similar structures perform similar functions and requires that structure is more conserved than sequence [198, 202]. Overall, the ARF Steward domains probably constitute a new family structurally related to Royal family, specific for plants, with no sequence similarities with other domains. *In vivo* research should clarify if the findings presented here are relevant *in planta*.

6.4.1 Steward domains as substitutes for histone PTMs antibodies

If it can be confirmed that AtARF1-DBD has specific histone-binding properties, this domain could be used as a tool in chromatin research. The characterization of histone marks requires clear assignment of locus specific modifications, as different PTMs are related with different interactors and responses. For these assignments, antibodies are the only tool for analysis. However, antibodies are usually not thoroughly tested for target specificity or cross reactivity, which is a major concern as it leads to inconsistencies in the interpretation of the results [187, 203, 204, 205]. In addition, antibodies suffer from certain biases in PTM recognition depending on the surrounding residues. For some histone readers, the interaction is strongest for tandem PTM modifications of the histone. An antibody would need to be raised [203] against this type of combined modifications to be able to detect their presence.

Another disadvantage of using antibodies is cumbersome production, which depends on animal immunization and later antibody isolation. This results in high costs and higher variability than with proteins produced in bacteria [204]. An initiative for using histone-interacting domains was previously proposed, but the production of a wide range of readers against all the different sequences requires a lot of effort due to the limited number of histone-interacting domains [204]. Protein engineering was proposed to overcome this limitation, but this would require a lot of trial and error. The presence of a functional histone-interacting domain in all ARFs of the auxin response family provides a tool for recognition of histone tails that cannot be detected using the known RF domains. The fact that several variations occur in the ARF residues surrounding the HC, potentially allows fine tuning recognition sequence by each ARF. As ARFs are ubiquitously found in plants for auxin response and there is a high number of ARFs encoded on higher plant genomes, there are a huge amount of potentially different PTM and/or histone variant specificities. The downside of using ARFs for chromatin research is that it will require years to analyse the specificities of different ARFs, as this is unexplored territory. On the other hand, the discovery of the presence of a domain capable of reading histone posttranslational modifications will hopefully attract interest in characterizing the ARF specificities in more detail. This will shed light on the ARF gene regulation mechanisms in certain cellular contexts. It will also allow the accumulation of information on the specificities of the different ARFs which has the potential to contribute to chromatin research.



Final considerations

| | | |
|-----------|--|------------|
| 7 | Discussion, conclusions and future perspectives | 111 |
| 7.1 | General Discussion | |
| 7.2 | Conclusions | |
| 7.3 | Future perspectives | |
| 8 | Materials and Methods | 119 |
| 8.1 | Protein expression and purification | |
| 8.2 | Preparation of dsDNA | |
| 8.3 | Crystallography | |
| 8.4 | RMSD values and distance calculation | |
| 8.5 | Analytical SEC | |
| 8.6 | SAXS analysis | |
| 8.7 | Dot-Blot assays | |
| 8.8 | Structural superposition analysis | |
| 8.9 | Peptide microarray assays | |
| 8.10 | Fluorescence anisotropy assays | |
| 8.11 | <i>Marchantia polymorpha</i> plant extracts | |
| 8.12 | Pull-down assays | |
| 8.13 | Overlay blots | |
| 9 | Supplementary information | 127 |
| 10 | Bibliography | 137 |
| | Publications | 155 |
| | Deposited structures (PDBs) | 156 |



7. Discussion, conclusions and future perspectives

7.1 General Discussion

Auxins are plant hormones involved in the regulation of many critical cellular processes. Among those processes, gene transcription regulation is one of the most studied. The main auxin, indole-3-acetic acid, is a small, rather simple molecule, and it is not completely understood how it can control as many processes as it is associated with. The principal regulation pathway, Nuclear Auxin Pathway (NAP), is a short response pathway with only 3 families of proteins involved: the ARFs, the Aux/IAAs and the TRANSPORT INHIBITOR RESISTANT1/ AUXIN F-BOX (TIR1/AFB) auxin receptors. Most land plants contain multiple homologs for each of these families, which provide variability to auxin response. For example, in the case of *Arabidopsis*, 23 ARFs, 29 Aux/IAAs and 6 TIR1/AFB have been identified. This big array of proteins gives rise to a huge amount of combinations of elements in NAP. However, the final effectors, the ARFs, were found to share high homology in the DNA binding residues, preserved during ARF evolution. Thus, it is difficult to understand how ARFs can select specific genes. The ‘molecular calliper’ hypothesis was proposed based on the binding mechanism observed in the first AtARF-DBD crystallographic structures. Under this hypothesis, ARF dimerization works like a calliper, measuring the distance between two consecutive AuxREs in inverted repeat configuration and specificity for DNA sequences depends on configuration of the B3 domains in each type of ARF and the distance of the AuxREs. This model has been proposed for other proteins as well, although it has been pointed out that, in plant genomes, there are fewer AuxREs in inverted repeat configuration than single AuxREs, and that this mode of binding is again insufficient to explain the variability of responses to auxin signalling. The first part of this thesis investigates the molecular calliper hypothesis in more detail, further proving and refining the model, and aims to find alternative models of specific gene selection that could work in concert with the molecular calliper mechanism. In the second part of this work we propose a complementary mechanism of specific gene selection, based on the presence of a Royal Family like module within ARF-DBDs, to which we refer as the Steward domain, with histone PTM binding properties.

7.1.1 The B3 conformational freedom is an ancestral mechanism of gene selection

Results presented in chapter two show that the main structural difference between classes and evolutionary distant proteins is the relative conformation between subdomains of the DBD. The structures of the ARF show that all the ARF-DBDs are formed by three subdomains, where the relative position of B3 and DD domains could determine the binding affinity of ARFs. The first α helix observed in the ARF structures is probably an important determinant for the position of the B3. As we have shown, these domains seem to share a similar position in MpARF2 and AtARF5, adapting to DNA through an outward rotation. In contrast, AtARF1 shows a planar structure between DD and B3 domains in the DNA-bound structure. This conformational difference is likely the cause for the observed differences in affinity of the different ARFs for DNA sequence containing inverted AuxRE at different spacings, where AtARF1 seem to prefer spacings of 8 nucleotides, whereas MpARF2 and AtARF5 prefer spacings of 7 nucleotides. The changes in affinities between ARF classes would fine-tune the competition of the ARFs for a DNA sequence, depending on the relative affinities for these sites and the relative concentrations of the ARFs. This fine-tuning mechanism seems to also occur in the simpler organism *Marchantia polymorpha* [28] and in more complex organisms. Thus, the intersubdomain conformation confers DNA specificity and seems to be an ancestral strategy of gene selection ubiquitous in the plant kingdom. This is reinforced by the fact that the B3 DNA interacting amino acids, dimerization interface and overall structure are conserved between *Arabidopsis thaliana*, *Marchantia polymorpha* and, as shown more recently, in *Chlorokybus atmophyticus*. Furthermore, it has been shown that the only CaARF existing in this organism interacts with inverted and direct sequences consisting of two AuxRE sites but not with single AuxRE sites, as is the case for *Arabidopsis* and *Marchantia* ARFs. This suggests that CaARF also dimerizes upon binding DNA, indicating that ARF DNA binding preferences were established in basal charophytes and maintained during evolution [55]. For *Arabidopsis* early diverged organisms as *Chlorokybus atmophyticus* this strategy may be sufficient for appropriate selection of sets of genes in gene regulation, but in complex organisms, more homologs of each class exist and this is probably necessary for the variability observed in response to auxin stimulus [206].

Although the residues involved in DNA binding are mostly conserved, there are some key differences that could complement the molecular calliper hypothesis. An important residue for DNA binding, His136 (AtARF1 numbering), is mostly conserved in classes A and B but not in class C ARFs. Clearly, this would establish a difference in binding affinity between classes A/B and C.

We have also hypothesized on the effect of DNA methylation on ARF DNA binding. A methylation site is present in some of the AuxRE sequences, which are methylated in a high percentage of sites on the plant genome. Strikingly, the methylation site in the high affinity DNA molecule would occur in the DNA bases that interact with His136. *In vivo* and *in vitro* studies are required to assess the importance of AuxRE methylation in ARF-mediated gene regulation, where methylation of AuxREs could constitute a new source of gene regulation.

7.1.2 Solution studies agree with crystallography findings

In chapter 3, we performed solution experiments to check all the inferences mentioned in chapter 2, which were based on the crystal structures. SEC and fluorescence anisotropy confirm the preference of AtARF1 towards higher AuxRE spacing and validate the observation that AtARF5 and MpARF2 are more similar in DNA binding characteristics than in sequence. This observation implies that

probably the longer $\alpha 1$ helix in MpARF2 and AtARF5 compared to AtARF1 is allowing DD and B3 subdomains to accommodate a wider range of AuxRE spacing. K_D values determined by FA suggest that the binding would be stronger for spacings of 7 or 8 nucleotides, where AtARF1 has a preference for spacings of 8 base pairs, and MpARF2 and AtARF5 prefer spacings of 7 base pairs. This spacing of 7-8 nucleotides between inverted AuxRE repeats allows for a cooperative binding. These results, though, did not clarify whether dimerization of ARFs is required prior to inverted AuxRE binding or if the DNA binding induced dimerization. For this purpose, SAXS analyses were conducted in ApoARFs and DNA-ARF complexes. SAXS clearly showed that ARF-DBDs can homodimerize *in vitro* without the participation of any other ARF domain or the DNA, but the addition of DNA considerably increases the homodimerization rate of all the class A/B ARFs tested. Remarkably, the inability of class C to homodimerize was confirmed. Computational predictions based in sequence and molecular modelling suggested that MpARF3 contains an insertion of 80 amino acids in $\alpha 6$ helix, an essential helix for AtARF1 and AtARF5 dimerization. In all experiments in solution, the result was always similar: MpARF3 forms monomers in SEC and SAXS in apo form and also interacts with the DNA as a monomer. These results contrast with dot-blot experiments, which suggest that MpARF3, AtARF1 and AtARF5 are able to heterodimerize. Surprisingly, the dimerization interface of MpARF3 seems to interact with that of AtARF1 and AtARF5 but not with itself. Possibly, the correctly formed dimerization interface of AtARF1 and AtARF5 would compensate for the insertion on MpARF3, whereas these insertions prevents homodimerization of MpARF3. The observation that ARFs from *Arabidopsis* and *Marchantia* are able to heterodimerize highlights the fact that the dimerization interface is highly conserved during evolution. These experiments also showed that AtARF1 is less prone to heterodimerize, since higher concentrations are required to reach the dimerization levels observed for MpARF2 and AtARF5. In the case of AtARF5, a similar level of heterodimerization was found with AtARF1 and MpARF2. This indicates that AtARF5, an activating ARF, could interact with other ARFs in the nucleus, allowing for crosstalk for fine tuning of its output. AtARF1, however, would prefer to homodimerize and inhibit gene expression. As it has been pointed before, competition between different ARFs for the binding sites is a determinant of the outcome of ARF-mediated gene regulation. Our results suggest that heterodimerization between ARF-DBDs occurs, which could also alter the outcome. The DNA binding properties of heterodimers is something that should be further studied.

7.1.3 Similarity between ARF-DBD may be explained by their *in vivo* activity

Part of the work presented here has been published in peer reviewed journals in cooperation with our collaborators of the Weijers group in the Netherlands. In one of the studies [28] we suggest that the similarity within ARFs may indicate competitive binding to the DNA binding sites, where regulation may be produced by two competing transcription factors underlying auxin response in the minimal *M. polymorpha* system. Class A-ARFs are the only auxin-sensitive transcriptional regulators in *Marchantia polymorpha* which switch on gene expression in an auxin-dependent manner. On the other hand, class B-ARFs function independently of auxin and antagonize A-ARF by competing for target sites and by recruiting the TPL co-repressor. Thus, expression patterns of A-ARF and B-ARF create zones in *M. polymorpha* with different auxin sensitivity. It is intuitive how auxin response may be regulated by modulating the stoichiometry of an antagonistic ARF pair. Interestingly, a single pair exists in *M. polymorpha*, and this basic module was expanded in evolutionary more evolved species, by duplication of the genes involved in auxin response. Duplication of A- or B-ARFs would allow differences in relative stoichiometry between activators and repressors, and also allow duplicated genes to “escape” ancestral regulation, for example by an increase in Aux/IAA interaction in B-ARFs. Strikingly, this simple model is consistent

with most studies in the more complex *Physcomitrella* and *Arabidopsis* organisms, rationalizing genetic interactions, protein interactions, activity assays and ARF gene expression patterns, and may therefore represent a universal unit at the base of the complex auxin response networks.

7.1.4 ARF Steward domain as selector of ARF gene specificity

As we have mentioned, the DNA specificity of the B3 is highly conserved between ARFs from different classes and species that diverged several million years ago. The refinement of the molecular calliper hypothesis indicates that, while it can affect gene specificity, it would probably act synergistically with other gene expression selection mechanisms. During the course of this work, and thanks to the solution of the *Marchantia polymorpha* ARF2-DBD structures presented in chapter 2, we demonstrated that MpARF2-DBD also contains a Tudor-like domain in the 80 C-terminal DBD amino acids, which was hitherto called the Ancillary Domain. Tudor domains belong to the Royal Family of histone PTM readers, and the presence of such a module in ARFs would point to a new way of gene selection through interactions with histone tails.

As we show in chapter 4, this domain is occluded in all ARF structures determined so far. Thus, no histone PTM binding is presumably expected. However, the conservation of the AD in ARFs throughout evolution suggests that these domains are functional. A thorough investigation of the similarities of the Ancillary Domain and RF proteins was performed through structural alignments in order to establish whether similar, occluded RF proteins exist in RF proteins. When reviewing Royal Family members to clarify ARF-AD belonging to this family, we observed several inconsistencies in the classification of the members in databases and in bibliography. For the purpose of classifying the ARF AD, we reviewed the different members proposed in different databases as well in bibliography and browsed the structures in search of similarities and differences between each member. This work suggested that some changes in the classification of some of the subfamilies could benefit the consistency of the current classification and would help in the classification of new RF PTM readers.

This comparison of RF family members and ARFs confirmed that the overall structures were very similar, with C α RMSDs close to 1 Å. In addition, the hydrophobic cage required for histone PTM recognition in all Royal family members is present in ARF-ADs and is properly formed. Therefore, this domain could be in fact functional. This hydrophobic cage was also conserved in all ARFs from *Arabidopsis* and *Marchantia* and searches using the AtARF1 Ancillary Domain amino acid sequence in protein sequence databases resulted in hits in other ARFs, while no hits outside of the plant kingdom were found, which suggests that the ancillary domain is present in many other ARFs and that this domain is exclusive to plants. This led us to propose that the AD is a member of the Royal Family, exclusive to plants. We then wanted to analyse the implication of the HC blocking residue observed in AtARF1, AtARF5 and MpARF2. According to our sequence alignments, the variation of the AtARF1-AD blocking residue, R343, is limited to arginine and lysine residues in class A/B ARFs, while in class C this residue is mostly an asparagine. This highlights the tendency of this domain to interact with basic residues. Compared to other Royal Family members, we discovered that some structures presented similar locking mechanism of their HCs. Since these domains lacked *in vitro* activity, they were classified as inactive. However, we demonstrate in chapter 5 that the AtARF1-DBD is able to interact with modified histone peptides in microarrays, where arginine methylation and lysine acylation of H2A and H2B histones are the preferred binders of AtARF1-DBD. Assuming that the HC of the ancillary domain is responsible for this interaction, we were interested in investigating whether the HC can be released from this locking mechanism. We found that pH affects the binding affinity. Thus, FA assays at neutral and basic pHs showed

no interaction with any of the typical RF histone peptides binders. When decreasing the pH to 4, we found that AtARF1 interacts with the H3 74:84 peptide, which was also suggested to be a weak binder in the microarray experiments. The K_D of this interaction was $6.91 \pm 0.83 \mu\text{M}$. When changing to pH 5.5, the affinity of this interaction was weaker, as evidence by a drop in K_D to $38.73 \pm 9.55 \mu\text{M}$. We also found that at this pH, the maximum anisotropy signal was lower than that observed at pH 4. Although pH levels as low as 4 are presumably never achieved physiologically, these experiments suggest that the inhibition mechanism of ARF-AD recognition is subjected to regulation, possibly by other cofactors or ARF-AD PTM. Notably, numerous phosphorylation sites are located within the ARF-AD which could play a role in this regulation process.

The structural alignments of Royal Family members with ARF-DBDs described in chapter 4 revealed a second putative Royal Family-like fold, located in Dimerization Domain. This dimerization domain was originally described to contain a “Taco shape” [39] and contains all the structural elements of a Royal Family domain, except for the Hydrophobic Cage, which only contains two of the five HC residues. This region is packed against $\alpha 2$ and 6, which blocks the entrance to a putative HC. As $\alpha 6$ helix is essential for ARF dimerization, we believe that the function of this RF-like region could be related to the stabilization of the $\alpha 6$ helix position. Considering the existence of Royal Family members where two repeated domains are required for the PTM recognition function, it is possible that this RF-like DD region is similarly required for putative binding of the AD to histone tails. The presence of two consecutive RF-like domains in ARF DBD is similar to the configuration found for the extended Tudor- and chromodomains. ARFs can therefore be classified as a distinct family of the Royal Family superfamily. We coined the term “Steward domain” to refer to the combined RF-like domains found in the DD and AD of the ARFs. The insertion of the B3 domain to this Steward domain, would combine DNA binding and histone PTM reading, creating an ancestral ARF-DBD. Combinations to this ancestral ARF-DBD module with MR and PB1 domains would allow it to actuate in auxin signalling.

7.1.5 Model of signal integration by ARFs

The data presented in this work suggest several additional mechanisms of ARF-mediated gene regulation under auxin stimulus. To accommodate for these in the model of ARF function, we integrate all the different inputs that affect ARF function, both those that are already described in literature and those suggested by the work described in this thesis, into a model that produces a complex, coordinated response. Under this model, ARF would function as a bridge between AuxRE recognition and histone PTMs in the expression of genes under auxin control. ARF will bind to the AuxRE through its B3 domain and to histone PTM through the ancillary domain when these elements are available for interaction. Once bound, the ARFs recruit partners to bring them into the vicinity of their substrate, as for example the chromatin-remodelling subunits BRAHMA and SPLAYED, that forms complexes with AtARF5 [24, 48]. Initial studies have shown that ARF middle regions play an important role in the recruitment of BRAHMA or SPLAYED to promoters of auxin-responsive genes. This recruitment, in the setting of nucleosome-bound ARFs, will lead to a change in chromatin structure and alter the exposure of DNA regions, allowing changes in gene expression.

The ARF-DBD dimer is similar in size to the Nucleosome Core Particle (NCP) which allows for several scenarios for the interaction between the two Figure 7.1. On the one hand, one monomer of the ARF dimer can contact one nucleosome, while the other monomer could be interacting with the neighbour nucleosome. On the other hand, several ARF-DBD dimers could bind to different nucleosomes, which then form oligomers through their PB1. Binding to multiple nucleosomes

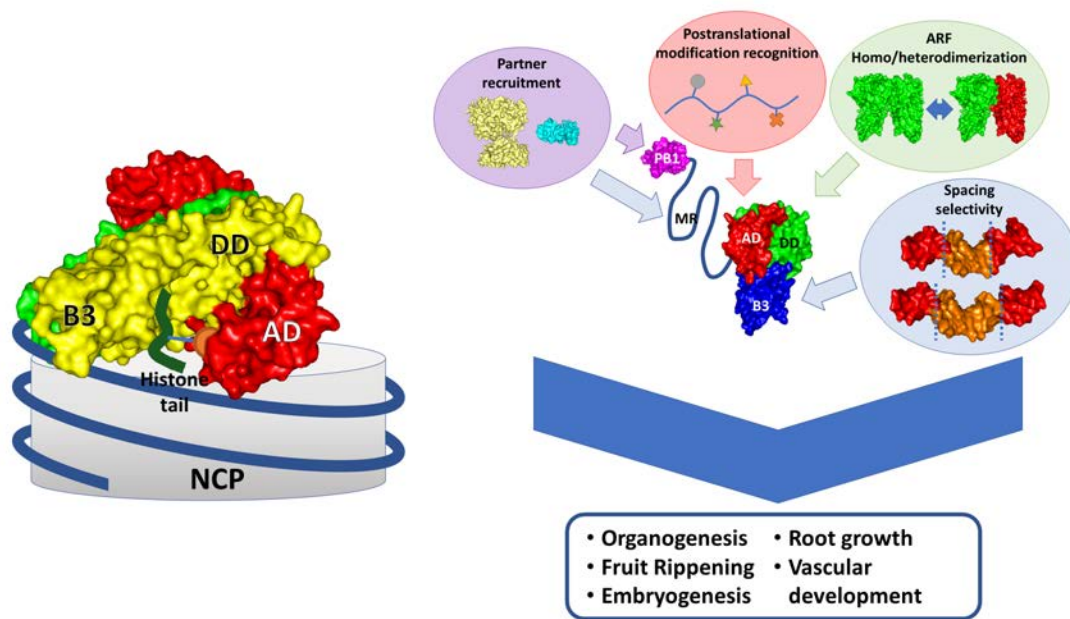


Figure 7.1: Model for the integration of information in auxin response. A: Scheme representing an ARF dimer (yellow and green) bound to exposed linker DNA in chromatin and histone tail posttranslational modifications (PTM). DNA is represented as a blue coiled string around a histone octamer (grey cylinder), which constitutes the nucleosome core particle (NCP). A histone tail is represented as a green curly line and a general PTM as an orange lollipop. The ancillary domains (AD, red). B: Schematic summary of the contribution of each ARF domain to the integration of the information from the cellular context on the coordinated response triggered by auxin. The combined information by ARFs is then translated into a composite cellular response to auxin signalling.

could recruit ARFs according to the affinities of the Steward domains for histones and associated PTMs, present in the distinct nucleosomes. As we have seen in previous chapters, the DNA sequence specificity and dimerization interface are mostly conserved between ARFs, even between evolutionary distant ARFs as MpARFs and AtARFs. The molecular calliper would also participate in the gene selection, as different spacing would in turn also determine the ARF binding to a gene promoter. ARF dimerization would combine affinities of ARF monomers for recognition sites, determined by the calliper model, and would integrate the effect of vicinal, distinct recognition sites from different nucleosomes. ARF heterodimerization would increase the complexity of the auxin response, where the affinity of the dimer depends both on the affinity of the different monomers for the recognition sites on the one hand, and affinity for the histone on the other. The resulting ARF homo or hetero dimer bound to the gene promoter will recruit partners given by the specificity determined by ARF Middle Region and PB1 domain. Those partners will produce the coordinated response for gene expression or repression, acting through chromatin modification, transcription machinery recruitment or other regulation mechanisms of gene expression.

In summary, this model integrates the contribution of the ARF calliper model of DNA recognition through the spatial arrangement of AuxREs with varying affinity for the ARFs; dimer variability due to ARF heterodimerization; integration of cellular status as outcome of other pathways resulting in epigenetic chromatin marks that ARFs can interpret; and the recruitment of different effectors to the AuxRE site. Although we detected affinity of ARFs for histone tail peptides, we cannot discard that the Ancillary Domain module could be involved in interactions with other proteins containing PTMs. This should be clarified in future studies of the ARF Ancillary Domain.

7.2 Conclusions

1. DBD architecture is conserved since *Marchantiophyta* and *Brassicaceae* divergence
2. α 1 helix length is determinant for the spacing tolerance of ARFs
3. H136 is essential to explain TGTCGG high affinity
4. ARF-DBDs are able to homodimerize in solution, although the main driving force is DNA interaction
5. 7-8 nucleotides AuxRE spacing is optimal for cooperative DNA binding
6. MpARF2 and AtARF5 prefer 7 nucleotides-spaced AuxRE while AtARF1 prefers 8 nucleotides
7. Heterodimerization is possible in solution, and the dimerization interface is sufficiently conserved to allow cross heterodimerization between distant species
8. RF family members have similar structural properties that have been summarized in this work. The analysis suggests that several families are very similar and should be merged into single groups
9. The Ancillary domain contains an entrance regulator that prevents the binding activity
10. The ARF-DD contains a secondary Tudor-like domain, whose contribution in AD function is unknown
11. ARF Ancillary domain is a new functional Royal Family member, exclusively present in plants, with the ability to recognize histone PTM
12. The Ancillary domain binding may require the dimerization domain for its function, similar to extended Tudor domains and extended chromodomains. This structural arrangement represents a new member of the Royal family, for which we propose the name “Steward domain”

7.3 Future perspectives

In this work we have found that DNA methylation, which has not been considered in ARF gene expression regulation, could have an impact in ARF gene regulation. Future studies should clarify the implication of AuxRE methylation on ARF signalling, and the implications *in vivo*.

To demonstrate the importance of Steward domains for ARF activity future research will be conducted on ARFs. For this purpose, we will introduce point mutations of critical ARF-AD residues, with the aim of identifying whether the HC residues induce changes in ARF signalling. Also, we will replace entire ADs with ADs from antagonist ARFs *in planta*, where we expect that those changes will change the functionality of the modified ARF, which can be monitored *in vivo*.

Given the implication of ARFs in gene selection, they represent a target for developing chemicals for weed control. At this moment, no herbicides targeting ARFs exist, so the development of new herbicides targeting ARFs would represent a ground-breaker on weed control, as no resistant populations would be present. This implies a need to study ARF Steward domains from different plants, especially non-desirable weeds, to develop specific AD blockers to prevent their growth. The analysis of these Steward domains will require huge efforts. As a proof-of-principle, we will continue to study the Steward domains of *Arabidopsis* ARFs.



8. Materials and Methods

8.1 Protein expression and purification

AtARFs-DBD, MpARFs-DBD and MpARF2-AD were expressed and purified using standard protocols as previously reported [39]. In brief, *E. coli* BL21 (DE3) Rosetta 2 were transformed with pTWIN1 vectors carrying the inserts, gently provided by Weijer's group. *E. coli* cells were then inoculated in fresh LB media complemented with $100\mu\text{g mL}^{-1}$ Ampicillin and $10\mu\text{g mL}^{-1}$ Chloramphenicol and grown at 37°C overnight. After that, the cells of the overnight culture were collected by centrifugation and suspended in the expression media (typically 2L of Terrific Broth complemented with $100\mu\text{g mL}^{-1}$ Ampicillin and $10\mu\text{g mL}^{-1}$ Chloramphenicol), at a ratio of 15 mL of preculture per liter of expression media. Cells in expression media were grown at 37°C until $O.D._{600} = 0.8 - 1$ was reached. After that, protein expression was induced overnight at 20°C by the addition of 0.3 mM isopropyl β -D-1- thiogalactopyranoside (IPTG, Omnipur).

After overnight induction, cells were collected by centrifugation at $4000\times g$, 30 minutes. Pellets were frozen and resuspended in lysis buffer (50 mM Hepes pH8, 1 mM EDTA, 500 mM NaCl, 2 mM MgCl_2 , 0.1% (v/v) Tween20) at a ratio of 2ml/g of cells. Typically, 10g of cells were obtained per liter of medium. The cell suspension was complemented with DNaseI ($10\mu\text{g mL}^{-1}$) and a tablet of cComplete™ EDTA-free Protease Inhibitor Cocktail (Roche, Cat. No. 05 056 489 001) and the suspension was sonicated for a total of 5 minutes, with cycles of 10 seconds on and 20 seconds off on an ice-water container. Insoluble matter was precipitated by centrifugation ($18000\times g$, 30 minutes). Supernatant was filtered through a $0.22\mu\text{m}$ filter and applied to a Chitin resin column (New England Biolabs) equilibrated in column buffer (20 mM Tris pH8, 50 mM NaCl, 0.1% (v/v) Tween20). The column was washed with 10 column volumes of column buffer. CBP tag was cut and ARFs were eluted by quickly flushing the column with 5 column volumes of column buffer supplemented with 50 mM DTT. After all the volume was passed through the column, the flow was stopped for 2 hours and let incubating at room temperature. ARFs were eluted flushing the column with column buffer. All the volume was concentrated with an Amicon ultra 10kDa MWCO (Millipore) and the buffer was exchanged using a PD-10 desalting column (GE Healthcare life

Table 8.1: List of annealed oligonucleotides sequences used in this work

| ID | Sequence |
|--------|-----------------------------|
| LFY | T TGTCAA TTTCCCAGC AAGACA A |
| TMO3 | T GGTCAA AAGTAAGAC TGGACC A |
| TMO5 | T GGTCTC TGGTCGG TCGACA A |
| TMO5Δ1 | T GGTCTC TGGTCGG TTTTTT A |
| TMO5Δ2 | T TTTTTT TGGTCGG TTTTTT A |
| 21-5 | T TGTCGG CATTG CCGACA A |
| 21-6 | T TGTCGG CGATCG CCGACA A |
| 21-7 | T TGTCGG CGATTTCG CCGACA A |
| 21-8 | T TGTCGG CGATATCG CCGACA A |
| 21-9 | T TGTCGG CGATTATCG CCGACA A |
| DR5 | T TGTCTC CCTTT TGTCTC A |
| ER7 | T TGTCTC CCAAAGG GAGACA A |
| ER8 | T TGTCTC CCTTTTGG GAGACA A |

sciences) to 20 mM Tris, 50 mM NaCl, pH 8 buffer. Protein was then further purified to homogeneity (to remove dimers and aggregates) by size exclusion chromatography with a Generon ProteoSEC 3-70, 16-600mm, in 20 mM Tris pH8, 50 mM NaCl. The chromatography was run at 1 ml/min on an ÄKTA Pure 25M and 280 and 260nm absorbance was recorded. Fractions containing the peak corresponding to ARF molecular weight were concentrated and filtered through 0.2µm. Purity was assessed to be > 95% by SDS-PAGE, followed by Coomassie Blue staining. Final protein concentration was determined by 280nm absorbance on a Nanodrop spectrophotometer (Thermo Scientific) and was used for assays immediately where possible or stored in aliquots at -80 °C.

8.2 Preparation of dsDNA

All DNAs were obtained from Biomers (Ulm, Germany). Forward and reverse complementary chains were diluted to 10 mM concentration in H_2O . The concentration of each chain was quantified by Abs(260nm), using the extinction coefficient calculated using OligoCalc (available at <http://biotools.nubic.northwestern.edu/OligoCalc.html>) for the single stranded oligo sequence. Then, equimolar amounts of forward and reverse chains were mixed and heated to >95°C for 5 minutes in a water bath. Then, the heating was stopped, and the bath was let slowly cool for 16h until reaching room temperature. The final concentration of the annealed sample was measured by Abs(260nm), using the extinction coefficient calculated using OligoCalc for the double stranded oligo sequence. The annealed sequences were stored at -20°C. A list of annealed oligonucleotide sequences tested in this work can be found in Table 8.1.

8.3 Crystallography

8.3.1 MpARF2-DBD:21ds C2

Purified MpARF2-DBD was concentrated and was mixed with an annealed 21ds dsDNA of sequence 5'-d(TTGTCGG CGATTTCG CCGACAA)-3' at a ratio of 2:1 (protein:DNA), to a final protein concentration of 5.3 mg/ml in 20 mM Tris (pH8.0), 500 mM NaCl, 40 mM DTT. Crystals

of MpARF2-DBD:21ds giving the highest resolution were obtained by the sitting-drop vapour-diffusion method at 18°C, by equilibration of drops of 1 µL protein + 1 µL crystallization buffer (0.2 M Sodium formate, 20% w/v Polyethylene glycol 3,350) against 100 µL of the crystallization buffer. Crystals grew to their maximum size in 3 days. Cryo-cooling in liquid nitrogen was performed by soaking crystals on a cryo-protecting solution consisting in reservoir complemented with 10% glycerol, followed by direct plunge-freezing in $N_2(L)$. Data collection was performed to the indicated resolutions at ALBA synchrotron Light Source on the BL13-Xaloc beamline [207]. The crystals belonged to space group $C2$, with one protein dimer and one dsDNA in the asymmetric unit. Data were processed with XDSgui [208]. See Supplementary Table 9.1 for further statistics.

8.3.2 MpARF2-DBD:21ds $I2_12_12_1$

Purified MpARF2-DBD was concentrated and was mixed with an annealed 21ds dsDNA of sequence 5'-d(TTGTCGG CGATTCG CCGACAA)-3' at a ratio of 2:1 (protein:DNA), to a final protein concentration of 5 mg/ml in 20 mM Tris (pH8.0), 500 mM NaCl, 20 mM DTT. Crystals of MpARF2-DBD:21ds giving the highest resolution were obtained by the sitting-drop vapour-diffusion method at 18°C, by equilibration of drops of 1 µL protein + 1 µL crystallization buffer (100 mM Tris pH 7, 27% PEG2KMME, 18.5% Glycerol) against 100 µL of the crystallization buffer. Cube-shaped crystals took one week to grow to the maximum size. Cryo-cooling in liquid nitrogen did not require using a cryo-protecting solution and was performed by direct plunge-freezing in $N_2(L)$. Data collection was performed to the indicated resolutions at ALBA synchrotron Light Source on the BL13-Xaloc beamline [207]. The crystals belonged to space group $I2_12_12_1$, with one protein molecule and one DNA recognition sequence in the asymmetric unit. Data were processed with XDSgui [208]. See Supplementary Table 9.4 for further statistics.

8.3.3 MpARF2-DBD:ER7 $I2_12_12_1$

Purified MpARF2-DBD was concentrated and was mixed with an annealed 21ds dsDNA of sequence 5'-d(TTGTCCTC CCTTTGG GAGACAA)-3' at a ratio of 2:1 (protein:DNA), to a final protein concentration of 4.6 mg/ml in 15 mM sodium HEPES (pH7.5), 150 mM NaCl, 20 mM DTT. Crystals of MpARF2-DBD:ER7 giving the highest resolution were obtained by the sitting-drop vapour-diffusion method at 18°C, by equilibration of drops of 1 µL protein + 1 µL crystallization buffer (0.2 M Ammonium sulphate, Tris pH 7, 0.1 M MES pH6.5, 24% PEG 3350) against 100 µL of the crystallization buffer. Cube-shaped crystals took one week to grow to the maximum size. Cryo-cooling in liquid nitrogen did not require using a cryo-protecting solution and was performed by direct plunge-freezing in $N_2(L)$. Data collection was performed to the indicated resolutions at ALBA synchrotron Light Source on the BL13-Xaloc beamline [207]. The crystals belonged to space group $I2_12_12_1$, with one protein molecule and one DNA recognition sequence in the asymmetric unit. Data were processed with XDSgui [208]. See Supplementary Table 9.3 for further statistics.

8.3.4 AtARF1-DBD:21ds $P2_1$

The AtARF1-DBD:21ds complex was prepared by mixing purified AtARF1-DBD in 20 mM Tris (pH8.0), 500 mM NaCl buffer with an annealed 21ds dsDNA of sequence 5'-d(TTGTCGG CCTTTGG CCGACAA)-3' in a 1:1 molar stoichiometry, to a final protein concentration of 20 mg/ml. An initial crystallization hit (B10 condition of the Morpheus HT screen, Molecular di-

mensions) was obtained using a full screen of the complex against sparse-matrix conditions. The diffraction data were obtained from crystals grown at 17 °C using the hanging drop vapour diffusion method, on 1:1 drops of the complex at 20 mg/ml protein concentration and crystallization buffer against 300 µl of crystallization buffer. The crystallization buffer corresponds to the B10 condition of the Morpheus screen (0.03 M Sodium fluoride; 0.03 M Sodium bromide; 0.03 M Sodium iodide, 0.1 M Tris-BICINE pH 8.5, 20% v/v Ethylene glycol; 10 % w/v PEG 8000). Crystals were frozen using Dual Thickness MicroLoops LD™ (Mitegen) by direct plunge freezing in liquid nitrogen. Data was collected in a single sweep at the XALOC beamline at the ALBA synchrotron [207]. The crystal belonged to space group $P2_1$. Data was processed using the Global phasing AutoPROC program [209]. The resolution was cut-off at 1.98 Å, see Supplementary Table 9.2 for further statistics.

8.3.5 Structure solution

The MpARF2-DBD:21ds $C2$ and AtARF1-DBD:21ds $P2_1$ structures were solved by molecular replacement with PHASER [210] from the CCP4 package [164] using the dimerization domain of AtARF1-DBD (PDB ID 4LDX). Near-complete initial models were obtained with Phenix autobuild [211]. The structures were completed through alternate manual model building with Coot v.0.8.9 [212] and refinement with PHENIX v.1.16-3549 [211]. The models were validated and further adjusted and refined using MolProbity [213]. The crystallographic and refinement parameters are given in Supplementary Table 9.1 and Table 9.2.

To solve the structure of MpARF2-DBD:21ds and MpARF2-DBD:ER7 (both at SG $I2_12_12_1$) we used the monomer of MpARF2-DBD:21ds (SG $C2$, PDBid: 6SDG) as template for molecular replacement with PHASER. In the case of AtARF1-DBD:21ds (SG $P2_1$), the monomer from AtARF1-DBD:ER7 (PDBid: 4LDX) was used for molecular replacement. The following refining steps were performed as indicated in the MpARF2-DBD:21ds $C2$ structure. See supplementary Table 9.4 and Table 9.3 for further details.

8.4 RMSD values and distance calculation

RMSD values and distances were calculated using PyMOL v.2.4.0a0 (www.pymol.org; Schrödinger LLC). All structure RMSD were calculated for all the $C\alpha$ of the corresponding chains using the PyMOL built-in align function. In the case of B3 RMSD calculation, the Dimerization Domain of the analysed ARFs were aligned as described for all structure. Then, the $C\alpha$ atoms of the B3 domains were selected and the RMSD values were computed using PyMOL built-in function `rms_cur`, providing the `matchmaker=-1` argument for proper calculations.

The distance measurements were done using the built-in PyMOL measure wizard for punctual distance measurements. In the case of multiple pairwise distance computation as for the measures taken in $\alpha5$ helix, a Python 3 adapted version of the script providing the `pairwise_dist` PyMOL function was written and used. The original script code can be found at https://pymolwiki.org/index.php/Pairwise_distances.

8.5 Analytical SEC

500 picomol of each tested dsDNA were diluted in SEC buffer (15 mM Hepes pH7.5, 150 mM NaCl) to a final volume of 25 μ L and were injected on a Superdex 200 increase 5/150 (GE Healthcare) equilibrated in SEC buffer. For protein binding assays, a molar ratio of 2:1 (protein:DNA) were set in test tubes and incubated on ice for 1h, maintaining as final buffer the SEC buffer. 25 μ L were injected in each case. For ApoARF control, the same amount of protein used for the assays was injected. The chromatography was run at 0.3ml/min on an ÄKTA Pure 25M, recording 280 and 260nm absorbance.

8.6 SAXS analysis

Different concentrations of AtARF1, AtARF5, MpARF2 and MpARF3 ranging from 0.7mg/ml to 7mg/ml were tested to record protein dimerization depending on concentration. For protein:DNA assays, a fixed protein concentration of 3.5mg/ml was used, while varying DNA concentration for 4:1, 2:1 and 1:1 DNA:protein stoichiometry. All the samples were prepared in a final buffer consisting of 20 mM Tris-HCl pH8.0, 150 mM NaCl, 1 mM DTT. SAXS data was collected at NCD-SWEET beamline (BL11, ALBA Synchrotron, Barcelona). The buffer and the buffer + DNA in all concentrations were collected for subtraction of protein samples. Measurements were carried out at 293 K in a quartz capillary of 1.5mm outer diameter and 0.01mm wall thickness. The data (20 frames with an exposure time of 0.5 sec/frame) was recorded using a Pilatus 1M detector (Dectris, Switzerland) at a sample-detector distance of 2.56 m and a wavelength of $\lambda = 1.0 \text{ \AA}$.

Buffer subtraction and extrapolation to infinite dilution were performed by using the program package primus/qt from the ATSAS 2.8.4 software suite [214]. The forward scattering $I(0)$ and the radius of gyration (R_g) were evaluated by using the Guinier approximation, and the maximum distance D_{max} of the particle was also computed from the entire scattering patterns with AutoGNOM. The excluded volume V_p of the particle was computed from the Porod invariant. The scattering from the crystallographic models was computed with CRY SOL [215]. The volume fractions of the oligomers were determined with OLIGOMER [76], using as probe the available PDB structures. In the case of MpARF3, the model generated in this work based on MpARF2 structure was used. *Ab initio* reconstructions of low resolution shapes were computed with GASBOR [77]. Crystallographic models were superposed to *ab initio* models with Supcomb [216]. The structure of the protein complex was refined to the GASBOR model by rigid body modeling by using the program Sreflex [217].

8.7 Dot-Blot assays

Dot-Blot assays were performed with unlabelled ARF samples in order to detect heterodimerization. PVDF membranes were activated under gentle agitation for 1' in MetOH followed by 2' H₂O and a final incubation in Tris-buffered saline supplemented with 0.1% (v/v) Tween-20 (TBST, 50 mM Tris-HCl, 150 mM NaCl, 0.1% (v/v) Tween-20, pH 7.6) for at least 5'. Then, a small quantity of TBST was left below the membrane to prevent membrane dehydration and the membrane surface was dried with a gentle stream of $N_2(g)$ before dispensing protein drops. The volume of the protein drops was of 1 μ L and between 5 and 0.04pmol protein in 1/2 dilutions were deposited in the membrane. The deposited samples were untagged AtARF1-DBD, AtARF5-DBD, MpARF2-DBD

and MpARF3-DBD. Lysozyme and a 6xHis-tagged protein were included as negative and positive controls, respectively.

Deposited drops were dried out under a gentle stream of $N_2(g)$ and TBST was immediately added to the membrane to prevent overdehydration. Then, the membrane was blocked with 5% skim milk in TBST for 1:30h and after this step, incubation with the labelled protein was done. The blocking, incubation and washing steps were performed at room temperature under agitation. AtARF1-DBD-5xHis or AtARF5-DBD-5xHis were diluted in 2.5% skim milk in TBST, at a final ARF concentration of $2\mu M$. A replicate membrane was incubated with 2.5% skim milk in TBST as control for antibody unspecific binding to the deposited samples. The incubation with the protein was done for 2:30h, followed by 5 washes for 5 minutes each with cold TBST. Mouse primary antibody (ThermoFisher Scientific Cat#MA1-135) was diluted 1:1000 in 2.5% skim milk TBST and membranes were incubated with the primary antibody solution under agitation for 1:15h, followed by 5x5' washes with cold TBST. As secondary antibody we used a Goat anti-Mouse IgG Fc Secondary Antibody, HRP conjugated (ThermoFisher Scientific, Cat#31437), following the incubation guidelines as for the primary with a 1:5000 antibody dilution. After the washing steps of the secondary antibody, the membrane was revealed using the HRP substrate SuperSignalTM West Pico PLUS Chemiluminescent Substrate (ThermoFisher Scientific Cat#34580) and the images were collected in the automatic exposure mode of the GE ImageQuant LAS 500. No unspecific antibody binding was detected in any of the assays.

8.8 Structural superposition analysis

Superposition of structures were performed the Superpose program available at the CCP4 suite. Default options were used. For multiple superpositions, a Python script was written to automate the structural superposition and analysis. The script performs the process in 3 steps: First, retrieves from the PDB the provided ids. Then, each structure retrieved from the PDB is superposed structurally to the reference structures. We used as reference the structures of the Ancillary Domains of AtARF1, AtARF5 and MpARF2 found in the PDBs 4LDX, 4LDU and 6SDG, respectively. Finally, the script collects the relevant information present in the superpose logs. The information analysed was the RMSD of the superposition, the amino acid sequence of the residues in the superposed structures sharing position with the HC residues and the blocking residue in the refence structure. The specific positions monitored were F298, F308, W335, E337, F342 and R343 for AtARF1-AD, F333, Y343, W370, E372, D377 and K378 for AtARF5-AD and F339, H349, W376, E378, E383 and R384 for MpARF2-AD. Each retrieved structure was assigned by the script to a RF subfamily based on the pfam database entry for each PDB id.

8.9 Peptide microarray assays

Microarray assays were based on previously reported studies [92, 218, 219]. Briefly, His-tagged proteins were diluted to $25\mu M$ in TBST and 5% (w/v) Bovine Serum Albumin (BSA, Sigma Cat#05479-50G) and incubated with Histone Code Peptide Microarray (JPT, Cat#His_MA_01 [179]) overnight at $4^\circ C$ in a humidity chamber. Arrays were washed three times, 3 minutes each with cold TBST and then probed with a mouse anti-6x-His Tag monoclonal antibody (ThermoFischer Scientific; Cat#MA1-135) freshly diluted to 1:1000 in TBST + 5% BSA. Arrays were washed again 3x with cold TBST and then probed with a Cy5-conjugated goat anti-mouse antibody at 1:5000 (Quimigen, Cat#610-110-121). Arrays were incubated with secondary antibody in low light

conditions at room temperature for 1:15h. After the incubation, the arrays were washed 3x3min with cold TBST and excess salt was removed in 0.1xTBS. The array was dried with a gentle stream of $N_2(g)$.

Arrays were imaged using a Typhoon Scanner using a resolution of 10 μ m pixel size, with excitation and emission filters for Cy5 (649/666nm). Protein binding was determined by densitometry with ImageQuant and Microsoft Excel as previously described [92], autocorrecting for background by subtracting the local fluorescence spot edge average. The average signal intensity for each peptide was then normalized to the most intense binding within an array, and normalized binding was averaged for three replicates. Heat maps of relative binding were generated using excel, maintaining the normalization of all the peptides analysed for cross comparison.

8.10 Fluorescence anisotropy assays

Binding of the DNA and peptides to ARFs was assayed in a CLARIOstar plate reader (BMG Labtech) on OptiPlate-384 Black well plates (PerkinElmer) in 10 μ L final assay volume. Peptides used for fluorescence anisotropy measurements were synthesized N-terminally labelled with fluorescein and purified by ThermoFisher Scientific. DNAs used for fluorescence anisotropy measurements were synthesized labelled with fluorescein at the 5' end of the forward chain and purified by Biomers GmbH. DNAs were annealed with the corresponding unlabelled reverse chain as described in this Materials and Methods section. The buffer used for anisotropy assays was 20mM Tris pH7.5, 150mM NaCl, 1mM DTT and 0.01% Triton X-100, excepting in the cases where different pHs were assayed. All buffers were properly degassed under vacuum and oxygen was removed saturating with nitrogen to prevent methionine oxidation. The final assay ARF concentration was varied from 0.1 to 500 μ M and the labelled ligand concentration was 50nM. An excitation wavelength of 485 nm and an emission wavelength of 528 nm were used. The data was measured at 25 °C and corrected for background by subtracting the free-labelled peptide signal. Plates were read immediately after preparation and later at 4 and 24h post preparation, storing plates in the dark at 4°C. Plates were warmed at room temperature for 30 minutes before reanalysis. While no differences were detected in the first 4 hours post preparation, measures at 24h were not considered due to significant fluorescence intensity loss and data accuracy. All the data treatment was done as previously described [220] and the data was fitted to Hill equation (or Michaelis-Menten when Hill coefficient equalled 1) using Origin 2018 (OriginLab Corporation). χ^2 values were used as the criteria for selecting between fittings.

8.11 *Marchantia polymorpha* plant extracts

Marchantia polymorpha (Strain Tak-1) material was gently provided by Ana Caño group (Centre for Research in Agricultural Genomics, CRAG). 3g of *Marchantia polymorpha* were frozen in $N_2(L)$ and liquefied with a mortar and pestle, with the addition of 2mL of cold extraction buffer (1xPBS, 0.5% Triton X-100, 0.5mM PMSF, 1x tablet cOmpleteTM, EDTA-free Protease Inhibitor Cocktail (Roche, Cat. No. 05 056 489 001)), with and without 2% SDS. All the plant remainders in the tools used were recovered with more extraction buffer up to a final volume of plant lysate of 12ml. To this solution, 1g of AlO_3 beads was added and vortexed for 30 seconds. The vortexing was repeated after incubating the sample for one minute on ice. The vortexed sample was further lysed by sonicating for five minutes on an ice-water container, in 10 ON/20 OFF seconds cycles. The final lysate was clarified by centrifugation at 4°C for 15 minutes at 18000xg. The supernatant

was filtered through 0.2 μ m filter and frozen in $N_2(L)$ in aliquots and stored at -80°C .

8.12 Pull-down assays

200 μ L of resuspended magnetic His-Trap resin (Quimigen, Reference 42179.01) were washed 3x with 500 μ L extraction buffer. After the final wash, the beads were resuspended in a final 200 μ L volume. 100 μ L of purified MpARF2-AD at 0.16mg/ml in 20mM Tris pH9, 500mM NaCl were added to half of the bead volume. 100 μ L of 20mM Tris pH9, 500mM NaCl were added to the remaining 100 μ L of resuspended beads as control. The tubes were incubated in agitation for 2h at room temperature and after applying a magnet, the supernatant was stored for SDS-PAGE analysis. The beads were washed 3x with 1mL wash buffer (Extraction buffer without Complete inhibitor tablet). The beads were then resuspended with 500 μ L of *Marchantia polymorpha* extract without SDS. The tube was closed under a $N_2(g)$ stream and sealed with BemisTM ParafilmTM M to prevent protein oxidation (we observed green to brown shift of the solution without using this method). The sample was incubated under agitation overnight at 4°C . After the incubation, the beads were washed 3x with 1mL Wash buffer and resuspended in 40 μ L wash buffer. All the samples were analysed by 12% SDS-PAGE for binding.

8.13 Overlay blots

Serial dilutions of the *Marchantia polymorpha* extracts obtained with and without 2% SDS were loaded on 12% SDS-PAGE. After the gel run, proteins were transferred to a PVDF membrane for 1:30h at 100V following standard Western Blot protocols [221]. In the case of enriched Overlay Blots, the *Marchantia polymorpha* extracts were substituted by pull-down samples obtained as specified in the previous section. After protein transference to the membrane, unspecific binding sites were blocked for 1:15h with 5% skimmed milk in TBST. The blocked membranes were incubated with a solution of 1 μ g/ml purified MpARF2-AD in 2.5% skimmed milk in TTBS. The incubation was done ON at 4°C , under agitation. After the incubation, membranes were washed 5x5 minutes with cold TBST and the membrane was incubated with mouse anti-His tag (ThermoFisher Scientific Cat#MA1-135) as primary antibody diluted 1:1000 in 2.5% skim milk TBST. Then, membranes were incubated with goat anti-mouse secondary antibody conjugated with HRP (ThermoFisher Scientific Cat#31431) diluted 1:5000 in 2.5% skim milk TBST. In all the cases the incubations were done at room temperature for 1:15h and 5x5 minutes washed were done after each antibody. Finally, membranes were revealed using a chemiluminescent HRP substrate (SuperSignalTM West Pico PLUS Chemiluminescent Substrate, ThermoFisher Scientific Cat#34580) and the images were collected in the automatic exposure mode of the GE ImageQuant LAS 500.

9. Supplementary information

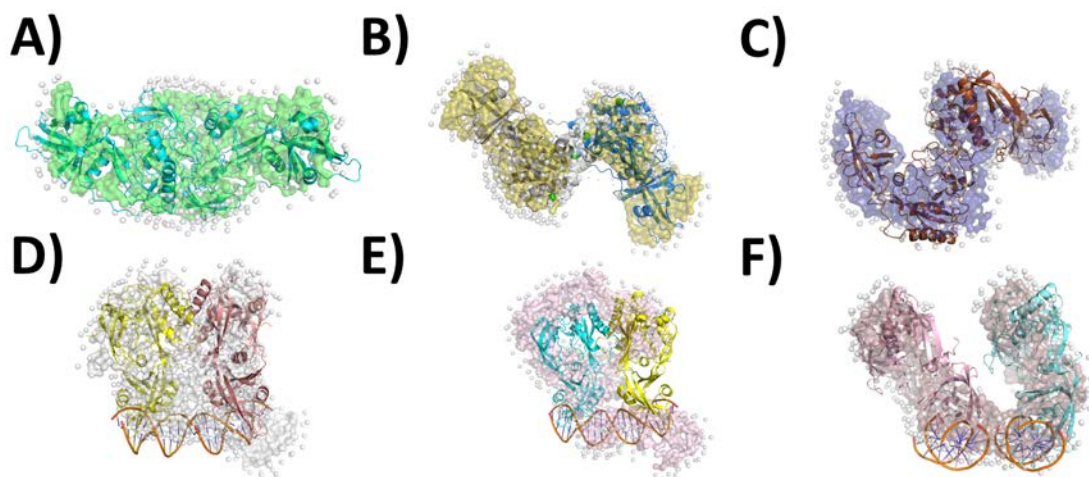


Figure 9.1: *Ab initio* Gashor models generated from SAXS data. Data from ApoAtARF1 (A), ApoAtARF5 (B) and ApoMpARF2 (C) was used to generate SAXS volumes, where crystallographic models were fitted in. SAXS data generated from complexes of 21ds with AtARF1 (D), AtARF5 (E) and MpARF2 (F) generated low quality models where crystallographic models were fitted in with difficulties.

Table 9.1: Summary of the data processing and refinement statistics of the crystallographic analysis of the 6SDG structure

| | | | |
|-----------------------------------|--|---|--------------|
| λ (Å) | 0.97919 | R_{cryst}^d / R_{free}^e (%) | 0.215, 0.278 |
| Space group | C2 | r.m.s. deviation from target values | |
| Unit cell parameters (Å, °) | a=162.77, b=79.36, c=79.77 $\alpha=90, \beta=116.93, \gamma=90$ | Bond lengths (Å) | 0.0132 |
| Resolution range (Å) ^a | 29.2- 2.96 (3.01-2.96) | Bond angles (°) | 1.451 |
| # of reflections: | | Molprobit scores | |
| total | 46775 (170) | Clashscore (‰) | 16.74 |
| unique | 14035 (51) | Poor rotamers (%) | 3.1 |
| Ellipsoidal Completeness (%) | 91.7 (30.6) | Ramachandran Outliers (%) | 0.81 |
| <I / Σ (I)> | 8.2 (0.7) | Ramachandran Favoured (%) | 93.37 |
| Average multiplicity | 3.3 (3.3) | Overall score (%) | 2.53 |
| R_{Rsym} (%) ^b | 8.9 | Isotropic B factor analysis | |
| R_{meas} (%) ^b | 10.6 | Average model B-factors (Å ²) | 112.0 |
| CC(1/2) (%) | 99.6 | B-factor from Wilson plot (Å ²) | 101.1 |

^a Throughout the table, the values in parentheses are for the outermost resolution shell.

^b $R_{Rsym} = \Sigma_h |\hat{I}_h - I_{h,i}| / \Sigma_h \Sigma_i I_{h,i}$, where $\hat{I}_h = (1/n_h) \Sigma_i I_{h,i}$ and n_h is the number of times a reflection is measured.

^c $R_{meas} = [\Sigma_h (n_h/[n_h-1])^{1/2} \Sigma_i |\hat{I}_h - I_{h,i}|] / \Sigma_h \Sigma_i I_{h,i}$, where $\hat{I}_h = (1/n_h) \Sigma_i I_{h,i}$ and n_h is the number of times a reflection is measured.

^d $R_{cryst} = \Sigma_{hkl} | |F_{obs}| - k |F_{calc}| | / \Sigma_{hkl} |F_{obs}|$

^e $R_{free} = \Sigma_{hkl \subset T} | |F_{obs}| - k |F_{calc}| | / \Sigma_{hkl \subset T} |F_{obs}|$ where T represents a test set comprising ~5% of all reflections excluded during refinement.

Table 9.2: Summary of the data processing and refinement statistics of the crystallographic analysis of the 6YCQ structure

| | | | |
|-----------------------------------|--|---|----------------|
| λ (Å) | 0.9793 | R_{cryst}^d / R_{free}^e (%) | 0.1717, 0.1991 |
| Space group | P21 | r.m.s. deviation from target values | |
| Unit cell parameters (Å, °) | a=43.3, b=102.78, c=127.04, $\beta=98.04$ | Bond lengths (Å) | 0.008 |
| Resolution range (Å) ^a | 47.57- 1.65 (1.68-1.65) | Bond angles (°) | 0.966 |
| # of reflections: | | Molprobit scores | |
| total | 277753 (269) | Clashscore (‰) | 6.88 |
| unique | 75146 (268) | Poor rotamers (%) | 0.16 |
| Ellipsoidal Completeness (%) | 88.1 (57.0) | Ramachandran Outliers (%) | 0.15 |
| <I / Σ (I)> | 15.9 (1.2) | Ramachandran Favoured (%) | 97.23 |
| Average multiplicity | 3.4 (3.3) | Overall score (%) | 1.52 |
| R_{Rsym} (%) ^b | 4.3 (80.5) | Isotropic B factor analysis | |
| R_{meas} (%) ^b | 5.2 (99.5) | Average model B-factors (Å ²) | 38.6 |
| CC(1/2) (%) | 99.9 (51.2) | B-factor from Wilson plot (Å ²) | 25.97 |

^a Throughout the table, the values in parentheses are for the outermost resolution shell.

^b $R_{Rsym} = \Sigma_h |\hat{I}_h - I_{h,i}| / \Sigma_h \Sigma_i I_{h,i}$, where $\hat{I}_h = (1/n_h) \Sigma_i I_{h,i}$ and n_h is the number of times a reflection is measured.

^c $R_{meas} = [\Sigma_h (n_h/[n_h-1])^{1/2} \Sigma_i |\hat{I}_h - I_{h,i}|] / \Sigma_h \Sigma_i I_{h,i}$, where $\hat{I}_h = (1/n_h) \Sigma_i I_{h,i}$ and n_h is the number of times a reflection is measured.

^d $R_{cryst} = \Sigma_{hkl} | |F_{obs}| - k |F_{calc}| | / \Sigma_{hkl} |F_{obs}|$

^e $R_{free} = \Sigma_{hkl \subset T} | |F_{obs}| - k |F_{calc}| | / \Sigma_{hkl \subset T} |F_{obs}|$ where T represents a test set comprising ~5% of all reflections excluded during refinement.

Table 9.3: Summary of the data processing and refinement statistics of the crystallographic analysis of the Mp2-ER7 structure

| | | | |
|-----------------------------------|--|---|----------------|
| λ (Å) | 1.0722 | R_{cryst}^d / R_{free}^e (%) | 0.1984, 0.2515 |
| Space group | $I2_12_12_1$ | r.m.s. deviation from target values | |
| Unit cell parameters (Å, °) | a=79.622, b=79.682, c=146.297, $\alpha = \beta = \gamma = 90$ | Bond lengths (Å) | 0.007 |
| Resolution range (Å) ^a | 73.148- 2.56 (2.74-2.56) | Bond angles (°) | 0.892 |
| # of reflections: | | Molprobit scores | |
| total | 156849 (7578) | Clashscore (‰) | 10.03 |
| unique | 12755 (638) | Poor rotamers (%) | 0 |
| Ellipsoidal Completeness (%) | 91.3 (41.8) | Ramachandran Outliers (%) | 0 |
| $\langle I / \Sigma(I) \rangle$ | 12.8 (1.0) | Ramachandran Favoured (%) | 96.07 |
| Average multiplicity | 12.3 | Overall score (%) | 1.79 |
| R_{Rsym} (%) ^b | 12.3 | Isotropic B factor analysis | |
| R_{meas} (%) ^b | 12.9 | Average model B-factors (Å ²) | 92.38 |
| CC(1/2) (%) | 99.5 (74.0) | B-factor from Wilson plot (Å ²) | 76.71 |

^a Throughout the table, the values in parentheses are for the outermost resolution shell.

^b $R_{Rsym} = \Sigma_h |\hat{I}_h - I_{h,i}| / \Sigma_h \Sigma_i I_{h,i}$, where $\hat{I}_h = (1/n_h) \Sigma_i I_{h,i}$ and n_h is the number of times a reflection is measured.

^c $R_{meas} = [\Sigma_h (n_h / [n_h - 1])^{1/2} \Sigma_i |\hat{I}_h - I_{h,i}|] / \Sigma_h \Sigma_i I_{h,i}$, where $\hat{I}_h = (1/n_h) \Sigma_i I_{h,i}$ and n_h is the number of times a reflection is measured.

^d $R_{cryst} = \Sigma_{hkl} | |F_{obs}| - k |F_{calc}| | / \Sigma_{hkl} |F_{obs}|$

^e $R_{free} = \Sigma_{hkl \subset T} | |F_{obs}| - k |F_{calc}| | / \Sigma_{hkl \subset T} |F_{obs}|$ where T represents a test set comprising ~5% of all reflections excluded during refinement.

Table 9.4: Summary of the data processing and refinement statistics of the crystallographic analysis of the Mp2-21 structure

| | | | |
|-----------------------------------|---|---|----------------|
| λ (Å) | 0.97926 | R_{cryst}^d / R_{free}^e (%) | 0.2173, 0.2450 |
| Space group | $I2_12_12_1$ | r.m.s. deviation from target values | |
| Unit cell parameters (Å, °) | a=79.911, b=80.933, c=, 146.623 $\alpha = \beta = \gamma = 90$ | Bond lengths (Å) | 0.005 |
| Resolution range (Å) ^a | 70.855 - 2.566 (2.610 – 2.566) | Bond angles (°) | 0.810 |
| # of reflections: | | Molprobit scores | |
| total | 150907 (7794) | Clashscore (‰) | 7.31 |
| unique | 15573 (764) | Poor rotamers (%) | 0 |
| Ellipsoidal Completeness (%) | 99.8 (99.7) | Ramachandran Outliers (%) | 0 |
| $\langle I / \Sigma(I) \rangle$ | 20.0 (1.5) | Ramachandran Favoured (%) | 95.67 |
| Average multiplicity | 9.7 | Overall score (%) | 1.70 |
| R_{Rsym} (%) ^b | 6.3 | Isotropic B factor analysis | |
| R_{meas} (%) ^b | 6.6 | Average model B-factors (Å ²) | 91.74 |
| CC(1/2) (%) | 99.9 | B-factor from Wilson plot (Å ²) | 78.72 |

^a Throughout the table, the values in parentheses are for the outermost resolution shell.

^b $R_{Rsym} = \Sigma_h |\hat{I}_h - I_{h,i}| / \Sigma_h \Sigma_i I_{h,i}$, where $\hat{I}_h = (1/n_h) \Sigma_i I_{h,i}$ and n_h is the number of times a reflection is measured.

^c $R_{meas} = [\Sigma_h (n_h / [n_h - 1])^{1/2} \Sigma_i |\hat{I}_h - I_{h,i}|] / \Sigma_h \Sigma_i I_{h,i}$, where $\hat{I}_h = (1/n_h) \Sigma_i I_{h,i}$ and n_h is the number of times a reflection is measured.

^d $R_{cryst} = \Sigma_{hkl} | |F_{obs}| - k |F_{calc}| | / \Sigma_{hkl} |F_{obs}|$

^e $R_{free} = \Sigma_{hkl \subset T} | |F_{obs}| - k |F_{calc}| | / \Sigma_{hkl \subset T} |F_{obs}|$ where T represents a test set comprising ~5% of all reflections excluded during refinement.

Table 9.5: Oligomer calculations for ApoARFs

| ARF | $[ARF](mg/ml)$ | Dimer | Monomer | Chi^2 |
|--------|----------------|-------------|-------------|---------|
| AtARF1 | 7 | 0.808±0.002 | 0.192±0.002 | 11.02 |
| | 3.5 | 0.601±0.003 | 0.399±0.003 | 1.77 |
| | 1.75 | 0.439±0.004 | 0.561±0.006 | 0.66 |
| | 0.7 | 0.229±0.012 | 0.771±0.017 | 0.51 |
| AtARF5 | 7 | 1.000±0.001 | 0.000±0.000 | 69.9 |
| | 3.5 | 0.966±0.003 | 0.034±0.004 | 11.6 |
| | 1.75 | 0.729±0.005 | 0.270±0.007 | 3.45 |
| | 0.7 | 0.576±0.011 | 0.424±0.016 | 0.9 |
| MpARF2 | 7 | 1.000±0.001 | 0.000±0.000 | 74.96 |
| | 3.5 | 1.000±0.001 | 0.000±0.000 | 16.98 |
| | 1.75 | 0.756±0.008 | 0.244±0.012 | 2.45 |
| | 0.7 | 0.393±0.018 | 0.606±0.027 | 0.75 |
| MpARF3 | 7 | 0.158±0.002 | 0.842±0.003 | 1.77 |
| | 3.5 | 0.062±0.004 | 0.938±0.006 | 0.68 |
| | 1.75 | 0.022±0.006 | 0.978±0.009 | 0.63 |
| | 0.7 | 0.000±0.000 | 1.000±0.008 | 0.9 |

Table 9.6: Oligomer calculations for ARF:DNA complexes

| ARF | DNA | Ratio | Dimer | Monomer | Chi^2 | ARF | DNA | Ratio | Dimer | Monomer | Chi^2 | |
|--------|------|-------|-------------|-------------|-------------|--------|-------|-------------|-------------|-------------|-------------|-------------|
| AtARF1 | 21-6 | 4 | 1.000±0.001 | 0.000±0.000 | 17.45 | MpARF2 | 21-6 | 4 | 0.918±0.004 | 0.082±0.006 | 17.26 | |
| | | 2 | 1.000±0.000 | 0.000±0.000 | 45.27 | | | 2 | 0.847±0.004 | 0.153±0.005 | 20.68 | |
| | | 1 | 1.000±0.001 | 0.000±0.000 | 18.92 | | | 1 | 0.699±0.003 | 0.301±0.004 | 28.56 | |
| | | 4 | 0.842±0.003 | 0.158±0.004 | 10.15 | | | 4 | 0.948±0.004 | 0.052±0.005 | 13.45 | |
| | 21-7 | 2 | 1.000±0.000 | 0.000±0.000 | 9.58 | | 21-7 | 2 | 0.960±0.003 | 0.040±0.004 | 12.55 | |
| | | 1 | 1.000±0.000 | 0.000±0.000 | 3.26 | | | 1 | 0.687±0.003 | 0.313±0.004 | 13.91 | |
| | | 4 | 0.835±0.003 | 0.165±0.003 | 10.63 | | | 4 | 0.920±0.004 | 0.080±0.005 | 13.61 | |
| | | 2 | 1.000±0.000 | 0.000±0.000 | 7.84 | | | 21-8 | 2 | 0.970±0.003 | 0.030±0.005 | 14.06 |
| | 21-8 | 1 | 1.000±0.000 | 0.000±0.000 | 2.59 | | 1 | | 0.665±0.003 | 0.335±0.004 | 15.57 | |
| | | 4 | 0.948±0.003 | 0.052±0.003 | 23.05 | | 4 | | 1.000±0.001 | 0.000±0.000 | 28.8 | |
| | | 21-9 | 2 | 1.000±0.000 | 0.000±0.000 | | 18.63 | | 21-9 | 2 | 0.952±0.003 | 0.048±0.005 |
| | | | 1 | 0.959±0.002 | 0.041±0.003 | | 15.58 | 1 | | 0.569±0.003 | 0.431±0.004 | 20.22 |
| AtARF5 | 21-6 | 4 | 1.000±0.001 | 0.000±0.000 | 29.29 | MpARF3 | 21-6 | 4 | 0.000±0.000 | 1.000±0.001 | 4.97 | |
| | | 2 | 1.000±0.001 | 0.000±0.000 | 33.55 | | | 2 | 0.000±0.000 | 1.000±0.001 | 5.82 | |
| | | 1 | 0.881±0.003 | 0.119±0.004 | 28.47 | | | 1 | 0.000±0.000 | 1.000±0.001 | 6.97 | |
| | | 4 | 1.000±0.001 | 0.000±0.000 | 24.64 | | | 4 | 0.000±0.000 | 1.000±0.001 | 6.41 | |
| | 21-7 | 2 | 1.000±0.001 | 0.000±0.000 | 18.33 | | 21-7 | 2 | 0.000±0.000 | 1.000±0.001 | 5.98 | |
| | | 1 | 0.920±0.003 | 0.080±0.004 | 24.56 | | | 1 | 0.000±0.000 | 1.000±0.001 | 10.98 | |
| | | 4 | 1.000±0.001 | 0.000±0.000 | 20.26 | | | 4 | 0.000±0.000 | 1.000±0.001 | 8.46 | |
| | | 21-8 | 2 | 1.000±0.001 | 0.000±0.000 | | | 20.32 | 21-8 | 2 | 0.000±0.000 | 1.000±0.001 |
| | 1 | | 0.922±0.003 | 0.078±0.003 | 23.58 | | 1 | 0.000±0.000 | | 1.000±0.001 | 13.59 | |
| | 4 | | 1.000±0.001 | 0.000±0.000 | 22.68 | | 4 | 0.000±0.000 | | 1.000±0.001 | 8.59 | |
| | 21-9 | | 2 | 1.000±0.001 | 0.000±0.000 | | 23.12 | 21-9 | | 2 | 0.000±0.000 | 1.000±0.001 |
| | | 1 | 0.769±0.003 | 0.231±0.003 | 21.98 | | 1 | | 0.000±0.000 | 1.000±0.001 | 14.44 | |

Table 9.7: SAXS parameters for Apo- and DNA-ARF complexes

| DNA | Ratio | AtARF1 | | | AtARF5 | | |
|------|-------|-----------|-------------------|---------------|----------------------|-------------------|---------------|
| | | Rg (nm) | $V_{porod}(nm^3)$ | $D_{max}(nm)$ | Rg(nm) | $V_{porod}(nm^3)$ | $D_{max}(nm)$ |
| 21-6 | 1 | 3.79±0.09 | 161.59 | 11.17 | 5.83±0.91 | 242.47 | 11.65 |
| | 2 | 3.76±0.99 | 178.15 | 11.18 | 4.15±0.34 | 209.02 | 12.56 |
| | 4 | 3.87±0.35 | 153.55 | 11.23 | 4.14±1.54 | 200.16 | 12.32 |
| 21-7 | 1 | 3.09±0.33 | 131.41 | 10.75 | 3.87±0.41 | 178.03 | 12 |
| | 2 | 3.56±0.21 | 135.56 | 10.49 | 3.85±0.19 | 188.83 | 11.64 |
| | 4 | 3.61±0.3 | 130.86 | 11.03 | 4.02±0.39 | 201.11 | 11.95 |
| 21-8 | 1 | 3.45±0.14 | 132.52 | 10.34 | 3.85±0.17 | 166.65 | 11.55 |
| | 2 | 4.42±0.55 | 142.99 | 9.87 | 3.99±0.14 | 198.88 | 12.07 |
| | 4 | 3.6±0.28 | 126.25 | 10.87 | 3.99±0.13 | 198.88 | 12.07 |
| 21-9 | 1 | 3.57±0.77 | 129.03 | 10.34 | 3.83±0.25 | 173.36 | 12.07 |
| | 2 | 3.72±0.17 | 144.5 | 11.25 | 4.24±0.49 | 228.69 | 12.83 |
| | 4 | 3.78±0.23 | 134.59 | 11.55 | 5.17±0.31 | 277.75 | 14.75 |
| Apo | 0.7 | 2.63±0.08 | 85.24 | 7.36 | 3.06±0.12 | 80.48 | 10.54 |
| | 1.75 | 2.92±0.02 | 83.58 | 7.6 | 3.34±0.34 | 84.64 | 8.52 |
| | 3.5 | 3.17±0.19 | 105.73 | 8.8 | 3.35±0.07 | 106.23 | 12.32 |
| | 7 | 3.48±0.09 | 111.84 | 10.89 | 3.89±0.39 | 153.94 | 11.96 |
| DNA | Ratio | MpARF2 | | | MpARF3 | | |
| | | Rg (nm) | $V_{porod}(nm^3)$ | $D_{max}(nm)$ | Rg(nm) | $V_{porod}(nm^3)$ | $D_{max}(nm)$ |
| 21-6 | 1 | 3.81±0.32 | 140.32 | 11.16 | 3.02±0.2 | 65.82 | 8.62 |
| | 2 | 3.88±0.18 | 159.74 | 11.24 | 3.06±0.03 | 67.45 | 8.42 |
| | 4 | 3.89±0.6 | 163.12 | 11.09 | 3.24±0.1 | 63.65 | 8.32 |
| 21-7 | 1 | 3.52±0.32 | 118.44 | 9.59 | 3.1±0.15 | 62.93 | 8.53 |
| | 2 | 3.49±0.32 | 139.9 | 9.84 | 3.07±0.15 | 68.04 | 8.52 |
| | 4 | 3.75±0.17 | 151.92 | 10.87 | 2.99±0.03 | 62.95 | 8.05 |
| 21-8 | 1 | 3.57±0.41 | 122.77 | 9.82 | 3.09±0.2 | 60.11 | 8.71 |
| | 2 | 3.64±0.35 | 146.35 | 9.93 | 3.14±0.09 | 65.49 | 8.82 |
| | 4 | 3.85±0.42 | 155.16 | 10.67 | 2.92±0.03 | 61.95 | 8.03 |
| 21-9 | 1 | 3.63±0.41 | 123.39 | 10.44 | 3.23±0.32 | 60.3 | 8.95 |
| | 2 | 3.76±0.42 | 163.16 | 11.01 | 3.06±0.17 | 64.8 | 8.87 |
| | 4 | 3.94±0.49 | 185.54 | 11.43 | 3.03±0.04 | 62 | 8.41 |
| Apo | 0.7 | 2.92±0.1 | 64.74 | 10.19 | Not enough intensity | | |
| | 1.75 | 3.31±0.08 | 76.71 | 7.87 | 2.49±0.09 | 53.5 | 6.48 |
| | 3.5 | 3.72±0.78 | 107.92 | 10.22 | 2.6±0.03 | 59.54 | 6.16 |
| | 7 | 6.52±0.1 | 181.18 | 11.01 | 2.77±0.1± | 59.87 | 7.34 |

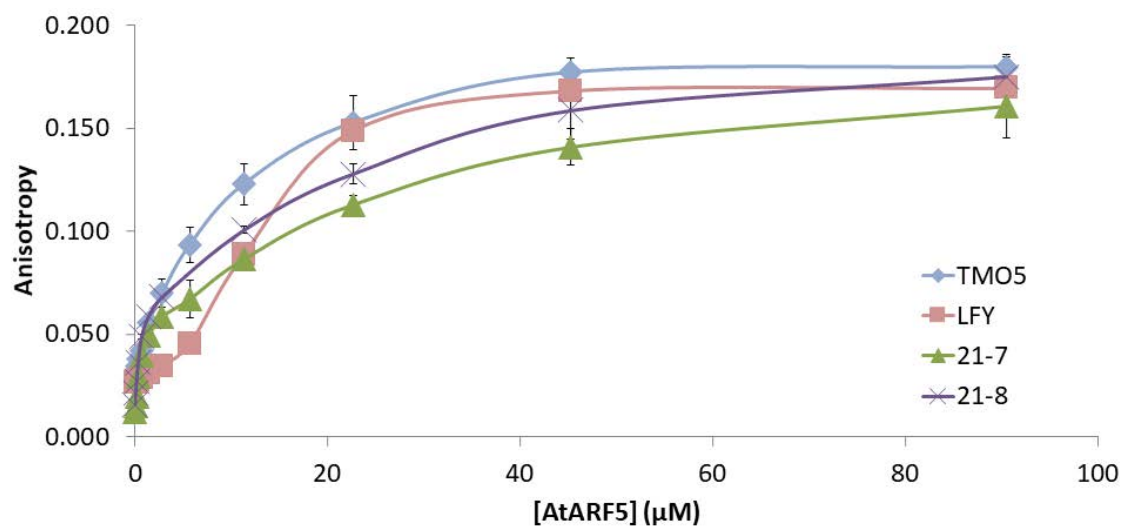


Figure 9.2: Anisotropy profile comparison. Sequences with an spacing comprised between seven and eight nucleotides result in a high interaction at low ARF concentrations, as depicted here by 21-7, 21-8 and TMO5 sequences. In contrast, the LFY sequence, with a spacing of nine nucleotides, shows a biphasic interaction with a small interaction at low concentrations.

Table 9.8: Chromo-like structures analysed

| Subfamily | PDB | UniprotID | pfam | SCOP | Interpro | CDD |
|----------------------|------|-----------|---------------|---------------|-----------------|---------------|
| Chromodomain | 2MJ8 | Q8N8U2 | Chromo | Chromo | Chromo | Chromo |
| | 1KNA | P05205 | Chromo | Chromo | Chromo | Chromo |
| | 2RNZ | Q08649 | Tudor-knot | - | RNA-knot chromo | Chromo barrel |
| | 2DY8 | P32657 | Chromo | Chromo | Chromo | Tandem Chromo |
| | 2EE1 | Q14839 | Chromo | Chromo | Chromo | Tandem Chromo |
| Chromo Barrel | 2K3Y | Q12432 | Tudor-knot | Chromo barrel | RNA-knot chromo | Chromo barrel |
| | 2LCC | P29374 | Tudor-knot | - | RNA-knot chromo | Chromo barrel |
| | 2BUD | O02193 | Tudor-knot | Chromo barrel | RNA-knot chromo | Chromo barrel |
| | 4QQG | Q92993 | Tudor-knot | - | RNA-knot chromo | Chromo barrel |
| | 4PL6 | Q4V3E2 | Tudor-knot | - | RNA-knot chromo | Chromo barrel |
| Chromo Shadow | 1E0B | P40381 | Chromo Shadow | Chromo | Chromo Shadow | Chromo Shadow |
| | 3P7J | P05205 | Chromo Shadow | - | Chromo Shadow | Chromo Shadow |
| | 3I3C | P45973 | Chromo Shadow | - | Chromo Shadow | Chromo Shadow |
| | 3KUP | Q13185 | Chromo Shadow | - | Chromo Shadow | Chromo Shadow |
| | 1DZ1 | P83917 | Chromo Shadow | Chromo | Chromo Shadow | Chromo Shadow |
| | 3Q6S | P83916 | Chromo Shadow | Chromo | Chromo Shadow | Chromo Shadow |

Table 9.9: Tudor-like structures analysed

| Subfamily | PDB | UniprotID | pfam | SCOP | Interpro | CDD |
|----------------|------|------------|-------------------------|------------|--------------|-------|
| Single Tudor | 4HCZ | O43189 | Tudor 2 | - | Tudor | - |
| | 3QII | Q9BVI0 | Tudor 3 | - | Tudor | Tudor |
| | 3P8D | Q9BVI0 | Tudor 3 | - | Tudor | Tudor |
| | 1MHN | Q16637 | SMN | Tudor | Tudor | Tudor |
| | 4A4E | Q16637 | SMN | Tudor | Tudor | Tudor |
| | 1G5V | Q16637 | SMN | Tudor | Tudor | Tudor |
| | 2LTO | Q9H7E2 | Tudor | - | Tudor | Tudor |
| | 5YJ8 | Q9H7E2 | Tudor | - | Tudor | Tudor |
| | 3PMT | Q9H7E2 | Tudor | - | Tudor | Tudor |
| | 3S6W | Q9H7E2 | Tudor | - | Tudor | Tudor |
| | 3PNW | Q9H7E2 | Tudor | - | Tudor | Tudor |
| | 4BD3 | Q5T6S3 | Tudor 2 | - | Tudor | Tudor |
| Tandem Tudor | 2LVM | Q12888 | Tudor | TP53bp1 | Tudor | Tudor |
| | 4H75 | Q9Y657 | Spin-Ssty | Spindlin-1 | Spindlin-1 | - |
| | 3ME9 | Q96ES7 | Tudor | - | Tudor | - |
| | 4TVR | Q96PU4 | Tandem Tudor/PHD-finger | - | Tandem Tudor | - |
| | 5YYA | Q96T88 | - | - | Tandem Tudor | - |
| Hybrid Tudor | 2QQS | O75164 | Tudor 2 | KDM4A | Tudor | Tudor |
| | 2XDP | Q9H3R0 | Tudor 2 | - | Tudor | - |
| | 4UC4 | O94953 | Tudor 2 | - | Tudor | - |
| Extended Tudor | 3OMC | Q7KZF4 | Tudor/SNase | SNase-like | SNase/Tudor | Tudor |
| | 2WAC | Q9W0S7 | Tudor | - | Tudor | Tudor |
| Plant Agenet | 5ZWX | A0A493R6M0 | - | - | Agenet | - |
| | 6IE4 | Q500V5 | - | - | Agenet | - |
| | 3H8Z | P51116 | Agenet/Tudor | - | Tudor | - |
| | 4OVA | Q06787 | Agenet/Tudor/KH_9 | - | Tudor | - |
| | 3KUF | P51114 | Agenet/Tudor | - | Tudor | - |

Table 9.10: Non-RF Histone methyllysine readers structures analysed

| Family | PDB | UniprotID | pfam | SCOP | Interpro | CDD |
|-------------------|------|-----------|------|----------|------------------------|---------------|
| MBT Domain | 2PQW | Q9Y468 | MBT | MBT | MBT | - |
| | 1OI1 | Q9UQR0 | MBT | Polycomb | MBT | - |
| | 2JTF | A8MW92 | MBT | - | Agenet/Tudor | Chromo barrel |
| | 2R57 | Q9VHA0 | MBT | - | MBT | - |
| | 1WJQ | Q96JM7 | MBT | - | MBT | - |
| PWWP | 5CIU | Q9UBC3 | PWWP | - | PWWP | PWWP |
| | 1H3Z | O94312 | PWWP | PWWP | IOC4-like/PWWP | PWWP |
| | 1KHC | O88509 | PWWP | PWWP | PWWP | PWWP |
| | 1N27 | Q9JMG7 | PWWP | PWWP | PWWP | PWWP |
| | 1RI0 | P51858 | PWWP | PWWP | PWWP | PWWP |
| PHD finger | 2G6Q | Q9ESK4 | - | ING2 | ING2/PHD domain | PHD finger |
| | 1F62 | Q9UIG0 | PHD | PHD | Zinc finger/PHD-finger | PHD finger |
| | 1FP0 | Q13263 | PHD | PHD | Zinc finger/PHD-finger | PHD finger |
| | 1MM2 | Q14839 | PHD | PHD | Zinc finger/PHD-finger | PHD |
| | 1WE9 | O81488 | PHD | PHD | Zinc finger/PHD-finger | PHD |
| WD40 | 2H13 | P61964 | WD40 | - | WD40 | WD40 |
| | 1A0R | P62871 | WD40 | WD40 | WD40 | WD40 |
| | 1ERJ | P16649 | WD40 | WD40 | WD40 | WD40 |
| | 1GG2 | P10824 | WD40 | WD40 | WD40 | WD40 |
| | 1NEX | P07834 | WD40 | WD40 | WD40 | WD40 |



10. Bibliography

1. Piras, V., Tomita, M. & Selvarajoo, K. Is central dogma a global property of cellular information flow? *Frontiers in Physiology* **3** NOV, 1–8. ISSN: 1664042X (2012) (cited on page 19).
2. Crick, F. Central Dogma of Molecular Biology. *Nature* **227**, 561–563. ISSN: 0028-0836. <https://doi.org/10.1038/227561a0> (Aug. 1970) (cited on page 19).
3. Stavreva, D. A., Varticovski, L. & Hager, G. L. Complex dynamics of transcription regulation. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* **1819**, 657–666. ISSN: 18749399. <http://dx.doi.org/10.1016/j.bbagr.2012.03.004> (2012) (cited on page 19).
4. Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237–1251. ISSN: 00928674. <http://dx.doi.org/10.1016/j.cell.2013.02.014> (2013) (cited on pages 19, 20).
5. Cramer, P. Organization and regulation of gene transcription. *Nature* **573**, 45–54. ISSN: 14764687. <http://dx.doi.org/10.1038/s41586-019-1517-4> (2019) (cited on page 19).
6. Leng, P. & Zhao, J. Transcription factors as molecular switches to regulate drought adaptation in maize. *Theoretical and Applied Genetics* **133**, 1455–1465. ISSN: 14322242. <https://doi.org/10.1007/s00122-019-03494-y> (2020) (cited on page 19).
7. Latchman, D. S. Transcription factors: An overview. *The International Journal of Biochemistry & Cell Biology* **29**, 1305–1312. ISSN: 13572725. <https://linkinghub.elsevier.com/retrieve/pii/S135727259700085X> (Dec. 1997) (cited on page 19).
8. Adachi, K. & Schöler, H. R. Directing reprogramming to pluripotency by transcription factors. *Current Opinion in Genetics and Development* **22**, 416–422. ISSN: 0959437X (2012) (cited on page 19).

9. Sikder, S., Kaypee, S. & Kundu, T. K. Regulation of epigenetic state by non-histone chromatin proteins and transcription factors: Implications in disease. *Journal of Biosciences* **45**, 15. ISSN: 0250-5991. <http://link.springer.com/10.1007/s12038-019-9974-3> (Dec. 2020) (cited on page 19).
10. Carlberg, C. & Molnár, F. *Human epigenomics* 1–217. ISBN: 9789811076145 (2018) (cited on page 19).
11. Robert, H. S. *et al.* Maternal auxin supply contributes to early embryo patterning in Arabidopsis. *Nature Plants* **4**, 548–553. ISSN: 20550278. <http://dx.doi.org/10.1038/s41477-018-0204-z> (2018) (cited on page 20).
12. Kam, R. K. T., Deng, Y., Chen, Y. & Zhao, H. Retinoic acid synthesis and functions in early embryonic development. *Cell and Bioscience* **2**, 11. ISSN: 20453701. <http://www.cellandbioscience.com/content/2/1/11> (2012) (cited on page 20).
13. Mironova, V., Teale, W., Shahriari, M., Dawson, J. & Palme, K. The Systems Biology of Auxin in Developing Embryos. *Trends in Plant Science* **22**, 225–235. ISSN: 13601385. <http://dx.doi.org/10.1016/j.tplants.2016.11.010> (2017) (cited on pages 20, 30).
14. Shaknovich, R. in *Advances in experimental medicine and biology* 133–150 (2013). ISBN: 978-1-4614-8050-1. <http://www.ncbi.nlm.nih.gov/pubmed/24014293> (cited on page 20).
15. Ballester, A. R. *et al.* Biochemical and molecular analysis of pink tomatoes: Deregulated expression of the gene encoding transcription factor SLMYB12 leads to pink tomato fruit color. *Plant Physiology* **152**, 71–84. ISSN: 00320889 (2010) (cited on page 20).
16. Huang, J., Li, Z. & Zhao, D. Deregulation of the OsmiR160 target gene OsARF18 causes growth and developmental defects with an alteration of auxin signaling in rice. *Scientific Reports* **6**, 1–14. ISSN: 20452322 (2016) (cited on page 20).
17. Sagor, G. H. M. *et al.* A novel strategy to produce sweeter tomato fruits with high sugar contents by fruit-specific expression of a single bZIP transcription factor gene. *Plant Biotechnology Journal* **14**, 1116–1126. ISSN: 14677644. <http://doi.wiley.com/10.1111/pbi.12480> (Apr. 2016) (cited on page 20).
18. UN Environment. *Global Environment Outlook – GEO-6: Healthy Planet, Healthy People* (ed UN Environment) 745. ISBN: 9781108627146. <https://www.unenvironment.org/resources/global-environment-outlook-6> (Cambridge University Press, May 2019) (cited on page 20).
19. Délye, C., Jasieniuk, M. & Le Corre, V. Deciphering the evolution of herbicide resistance in weeds. *Trends in Genetics* **29**, 649–658. ISSN: 01689525 (2013) (cited on page 20).
20. Quareshy, M., Prusinska, J., Li, J. & Napier, R. A cheminformatics review of auxins as herbicides. *Journal of Experimental Botany* **69**, 265–275. ISSN: 14602431 (2018) (cited on pages 20, 21).
21. Do, B. H., Phuong, V. T. B., Tran, G. B. & Nguyen, N. H. Emerging functions of chromatin modifications in auxin biosynthesis in response to environmental alterations. *Plant Growth Regulation* **87**, 165–174. ISSN: 15735087. <http://dx.doi.org/10.1007/s10725-018-0453-x> (2019) (cited on page 20).
22. Paque, S. & Weijers, D. Q&A: Auxin: the plant molecule that influences almost anything. *BMC Biology* **14**, 67. ISSN: 1741-7007. <http://bmcbiol.biomedcentral.com/articles/10.1186/s12915-016-0291-0> (2016) (cited on pages 20, 21).

23. Perrot-Rechenmann, C. Cellular Responses to Auxin: Division versus Expansion. *Cold Spring Harbor Perspectives in Biology* **2**, 1–15. ISSN: 1943-0264. <http://cshperspectives.cshlp.org/lookup/doi/10.1101/cshperspect.a001446> (May 2010) (cited on page 20).
24. Weijers, D. & Wagner, D. Transcriptional Responses to the Auxin Hormone. *Annual Review of Plant Biology* **67**, 539–574. ISSN: 1543-5008. <http://www.annualreviews.org/doi/10.1146/annurev-arplant-043015-112122> (2016) (cited on pages 20, 21, 28, 115).
25. Lavy, M. & Estelle, M. Mechanisms of auxin signaling. *Development* **143**, 3226–3229. ISSN: 0950-1991. <http://dev.biologists.org/lookup/doi/10.1242/dev.131870> (2016) (cited on pages 20, 21).
26. Kato, H., Nishihama, R., Weijers, D. & Kohchi, T. Evolution of nuclear auxin signaling: lessons from genetic studies with basal land plants. *Journal of Experimental Botany* **69**, 291–301. ISSN: 0022-0957. <http://academic.oup.com/jxb/article/doi/10.1093/jxb/erx267/4068714/Evolution-of-nuclear-auxin-signaling-lessons-from> (2017) (cited on pages 20, 22, 28, 58).
27. Chandler, J. W. Auxin response factors. *Plant Cell and Environment* **39**, 1014–1028. ISSN: 13653040 (2016) (cited on pages 20, 43, 45, 58, 93).
28. Kato, H. *et al.* Design principles of a minimal auxin response system. *Nature Plants* **6**, 473–482. ISSN: 2055-0278. <http://dx.doi.org/10.1038/s41477-020-0662-y> (May 2020) (cited on pages 20, 22, 27, 31, 33, 52, 112, 113).
29. Vernoux, T. *et al.* The auxin signalling network translates dynamic input into robust patterning at the shoot apex. *Molecular Systems Biology* **7**. ISSN: 17444292 (2011) (cited on pages 20, 44).
30. Nanao, M. H. *et al.* Structural basis for oligomerization of auxin transcriptional regulators. *Nature Communications* **5**, 3617. ISSN: 2041-1723. <http://www.nature.com/articles/ncomms4617> (May 2014) (cited on pages 21, 22).
31. Ulmasov, T., Hagen, G. & Guilfoyle, T. J. ARF1, a Transcription Factor That Binds to Auxin Response Elements. *Science* **276**, 1865–1868. ISSN: 00368075. <http://www.sciencemag.org/cgi/doi/10.1126/science.276.5320.1865> (June 1997) (cited on pages 21, 22).
32. Roosjen, M., Paque, S. & Weijers, D. Auxin Response Factors: output control in auxin biology. *Journal of Experimental Botany* **69**, 179–188. ISSN: 0022-0957. <https://academic.oup.com/jxb/article-lookup/doi/10.1093/jxb/erx237> (2017) (cited on pages 21, 28, 31, 44, 45, 58).
33. Li, C. *et al.* Concerted genomic targeting of H3K27 demethylase REF6 and chromatin-remodeling ATPase BRM in Arabidopsis. *Nature Genetics* **48**, 687–693. ISSN: 15461718. <http://dx.doi.org/10.1038/ng.3555> (2016) (cited on pages 21, 28, 31, 58, 86).
34. Dinesh, D. C. *et al.* Solution structure of the PsIAA4 oligomerization domain reveals interaction modes for transcription factors in early auxin response. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 6230–6235. ISSN: 10916490 (2015) (cited on pages 21, 22).
35. Guilfoyle, T. J. The PB1 Domain in Auxin Response Factor and Aux/IAA Proteins: A Versatile Protein Interaction Module in the Auxin Response. *The Plant Cell* **27**, 33–43. ISSN: 1040-4651. <http://www.plantcell.org/lookup/doi/10.1105/tpc.114.132753> (Jan. 2015) (cited on pages 21, 28, 31, 58).

36. Kim, Y. *et al.* Determinants of PB1 Domain Interactions in Auxin Response Factor ARF5 and Repressor IAA17. *Journal of Molecular Biology*, 1–13. ISSN: 10898638. <https://doi.org/10.1016/j.jmb.2020.04.007> (2020) (cited on page 21).
37. Korasick, D. A. *et al.* Molecular basis for AUXIN RESPONSE FACTOR protein interaction and the control of auxin response repression. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 5427–32 (2014) (cited on pages 21, 22, 28, 31, 95).
38. Stigliani, A. *et al.* Capturing auxin response factors syntax using DNA binding models. *Molecular Plant*, 1–11. ISSN: 16742052. <https://linkinghub.elsevier.com/retrieve/pii/S167420521830306X> (2018) (cited on pages 21, 28, 31, 58).
39. Boer, D. R. *et al.* Structural Basis for DNA Binding Specificity by the Auxin-Dependent ARF Transcription Factors. *Cell* **156**, 577–589. ISSN: 00928674. <http://linkinghub.elsevier.com/retrieve/pii/S0092867413015961> (2014) (cited on pages 21, 22, 28, 30, 31, 36, 39, 40, 44, 52, 58, 77–79, 84, 90, 93, 115, 119).
40. Guilfoyle, T. J. & Hagen, G. Auxin response factors. *Current Opinion in Plant Biology* **10**, 453–460. ISSN: 13695266. arXiv: NIHMS150003 (2007) (cited on pages 21, 86).
41. Odat, O. *et al.* Characterization of an allelic series in the MONOPTEROS gene of arabidopsis. *Genesis* **52**, 127–133. ISSN: 1526954X (2014) (cited on page 22).
42. Shen, C. *et al.* Functional analysis of the structural domain of ARF proteins in rice (*Oryza sativa* L.) *Journal of Experimental Botany* **61**, 3971–3981. ISSN: 1460-2431. <https://academic.oup.com/jxb/article-lookup/doi/10.1093/jxb/erq208> (Sept. 2010) (cited on page 22).
43. Mutte, S. K. *et al.* Origin and evolution of the nuclear auxin response system. *eLife* **7**, e33399. ISSN: 2050-084X. <https://elifesciences.org/articles/33399> (2018) (cited on pages 22, 28, 31, 58).
44. Kato, H. *et al.* Auxin-Mediated Transcriptional System with a Minimal Set of Components Is Critical for Morphogenesis through the Life Cycle in *Marchantia polymorpha*. *PLoS Genetics* **11**, 1–26. ISSN: 15537404 (2015) (cited on pages 22, 31).
45. Ulmasov, T., Hagen, G. & Guilfoyle, T. J. Activation and repression of transcription by auxin-response factors. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 5844–5849. ISSN: 0027-8424 (1999) (cited on pages 22, 44).
46. Nguyen, C. T., Tran, G.-B. & Nguyen, N. H. Homeostasis of histone acetylation is critical for auxin signaling and root morphogenesis. *Plant Molecular Biology* **103**, 1–7. ISSN: 0167-4412. <https://doi.org/10.1007/s11103-020-00985-1> (May 2020) (cited on pages 22, 28, 57, 58).
47. Tiwari, S. B., Hagen, G. & Guilfoyle, T. The Roles of Auxin Response Factor Domains in Auxin-Responsive Transcription. *The Plant Cell* **15**, 533–543. ISSN: 1664-462X. arXiv: 15334406. <https://doi.org/10.1105/tpc.008417> (Feb. 2003) (cited on page 22).
48. Wu, M. F. *et al.* Auxin-regulated chromatin switch directs acquisition of flower primordium founder fate. *eLife* **4**, 1–20. ISSN: 2050084X (2015) (cited on pages 22, 28, 58, 115).
49. Li, S.-B. *et al.* Genome-wide identification, isolation and expression analysis of auxin response factor (ARF) gene family in sweet orange (*Citrus sinensis*). *Frontiers in plant science* **6**, 119. ISSN: 1664-462X. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4378189> (2015) (cited on page 22).

50. Flores-Sandoval, E. *et al.* Class C ARFs evolved before the origin of land plants and antagonize differentiation and developmental transitions in Marchantia polymorpha. *New Phytologist* **218**, 1612–1630. ISSN: 14698137. <http://doi.wiley.com/10.1111/nph.15090> (2018) (cited on pages 22, 30).
51. Freire-Rios, A. *et al.* Architecture of DNA elements mediating ARF transcription factor binding and auxin-responsive gene expression in Arabidopsis. *Proceedings of the National Academy of Sciences*, 202009554. ISSN: 0027-8424. <http://www.pnas.org/lookup/doi/10.1073/pnas.2009554117> (Sept. 2020) (cited on pages 27, 31, 33, 39).
52. Finet, C., Berne-Dedieu, A., Scutt, C. P. & Marlétaz, F. Evolution of the ARF gene family in land plants: Old domains, new tricks. *Molecular Biology and Evolution* **30**, 45–56. ISSN: 07374038 (2013) (cited on page 28).
53. Kubeš, M. & Napier, R. Non-canonical auxin signalling: fast and curious. *Journal of Experimental Botany* **70**, 2609–2614. ISSN: 0022-0957 (2019) (cited on page 28).
54. Kim, J., Harter, K. & Theologis, A. Protein-protein interactions among the Aux/IAA proteins. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 11786–11791. ISSN: 00278424. <http://www.pnas.org/cgi/doi/10.1073/pnas.94.22.11786> (Oct. 1997) (cited on page 28).
55. Martin-Arevalillo, R. *et al.* Evolution of the Auxin Response Factors from charophyte ancestors. *PLOS Genetics* **15** (ed Reed, J.) e1008400. ISSN: 1553-7404. <http://dx.plos.org/10.1371/journal.pgen.1008400> (Sept. 2019) (cited on pages 28, 30, 48, 80, 82, 112).
56. Korasick, D. A., Jez, J. M. & Strader, L. C. Refining the nuclear auxin response pathway through structural biology. *Current Opinion in Plant Biology* **27**, 22–28. ISSN: 13695266. <http://dx.doi.org/10.1016/j.pbi.2015.05.007> (2015) (cited on pages 28, 31, 58).
57. Yamasaki, K., Kigawa, T., Seki, M., Shinozaki, K. & Yokoyama, S. DNA-binding domains of plant-specific transcription factors: Structure, function, and evolution. *Trends in Plant Science* **18**, 267–276. ISSN: 13601385. <http://dx.doi.org/10.1016/j.tplants.2012.09.001> (2013) (cited on pages 28, 29).
58. Yamasaki, K. *et al.* Solution structure of the B3 DNA binding domain of the Arabidopsis cold-responsive transcription factor RAV1. *Plant Cell* **16**, 3448–3459. ISSN: 10404651 (2004) (cited on pages 28, 29).
59. Waltner, J. K., Peterson, F. C., Lytle, B. L. & Volkman, B. F. Structure of the B3 domain from Arabidopsis thaliana protein At1g16640. *Protein Science* **14**, 2478–2483. ISSN: 09618368 (2005) (cited on pages 28, 32, 33).
60. Kagaya, Y., Ohmiya, K. & Hattori, T. RAV1, a novel DNA-binding protein, binds to bipartite recognition sequence through two distinct DNA-binding domains uniquely found in higher plants. *Nucleic Acids Research* **27**, 470–478. ISSN: 0305-1048. <http://www.sciencemag.org/cgi/doi/10.1126/science.1205687> (Jan. 1999) (cited on page 29).
61. Gehring, M. & Henikoff, S. DNA Methylation and Demethylation in Arabidopsis. *The Arabidopsis Book* **6**, e0102. ISSN: 1543-8120. <http://www.bioone.org/doi/abs/10.1199/tab.0102> (Jan. 2008) (cited on page 29).
62. Briollais, L. & Durrieu, G. Application of quantile regression to recent genetic and -omic studies. *Human Genetics* **133**, 951–966. ISSN: 0340-6717. <http://link.springer.com/10.1007/s00439-014-1440-6> (Aug. 2014) (cited on pages 29, 96).

63. Capuano, F., Mülleder, M., Kok, R., Blom, H. J. & Ralser, M. Cytosine DNA methylation is found in *Drosophila melanogaster* but absent in *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and other yeast species. *Analytical Chemistry* **86**, 3697–3702. ISSN: 15206882 (2014) (cited on pages 29, 41).
64. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322. ISSN: 0028-0836. <http://www.nature.com/articles/nature08514> (Nov. 2009) (cited on page 29).
65. Chen, W. F. *et al.* Structural analysis reveals a “molecular calipers” mechanism for a LATERAL ORGAN BOUNDARIES DOMAIN transcription factor protein from wheat. *Journal of Biological Chemistry* **294**, 142–156. ISSN: 1083351X (2019) (cited on page 31).
66. Schlereth, A. *et al.* MONOPTEROS controls embryonic root initiation by regulating a mobile transcription factor. *Nature* **464**, 913–916. ISSN: 00280836 (2010) (cited on page 31).
67. Yamaguchi, N. *et al.* A Molecular Framework for Auxin-Mediated Initiation of Flower Primordia. *Developmental Cell* **24**, 271–282. ISSN: 15345807. <http://dx.doi.org/10.1016/j.devcel.2012.12.017> (2013) (cited on page 31).
68. O’Malley, R. C. *et al.* Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* **165**, 1280–1292. ISSN: 10974172. <http://dx.doi.org/10.1016/j.cell.2016.04.038> (2016) (cited on page 38).
69. Siódmiak, J. *et al.* Molecular Dynamic Analysis of Hyaluronic Acid and Phospholipid Interaction in Tribological Surgical Adjuvant Design for Osteoarthritis. *Molecules* **22**, 1436. ISSN: 1420-3049. <http://www.mdpi.com/1420-3049/22/9/1436> (Sept. 2017) (cited on page 41).
70. Onofrio, A. *et al.* Distance-dependent hydrophobic–hydrophobic contacts in protein folding simulations. *Phys. Chem. Chem. Phys.* **16**, 18907–18917. ISSN: 1463-9076. <http://xlink.rsc.org/?DOI=C4CP01131G> (2014) (cited on page 41).
71. Amoutzias, G. D., Robertson, D. L., Van de Peer, Y. & Oliver, S. G. Choose your partners: dimerization in eukaryotic transcription factors. *Trends in Biochemical Sciences* **33**, 220–229. ISSN: 09680004 (2008) (cited on page 44).
72. Hardtke, C. S. *et al.* Overlapping and non-redundant functions of the Arabidopsis auxin response factors MONOPTEROS and NONPHOTOTROPIC HYPOCOTYL 4. *Development* **131**, 1089–1100. ISSN: 09501991 (2004) (cited on page 44).
73. Flores-Sandoval, E., Eklund, D. M. & Bowman, J. L. A Simple Auxin Transcriptional Response System Regulates Multiple Morphogenetic Processes in the Liverwort *Marchantia polymorpha*. *PLoS Genetics* **11**, 1–26. ISSN: 15537404 (2015) (cited on page 45).
74. Guilfoyle, T. J. & Hagen, G. Auxin response factors. *Journal of Plant Growth Regulation* **20**, 281–291. ISSN: 07217595 (2001) (cited on page 45).
75. Wang, S., Hagen, G. & Guilfoyle, T. J. ARF-Aux/IAA interactions through domain III/IV are not strictly required for auxin-responsive gene expression. *Plant Signaling and Behavior* **8**, e24526. ISSN: 15592316. <https://dx.doi.org/10.4161%2Fpsb.24526> (2013) (cited on page 45).
76. Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. & Svergun, D. I. PRIMUS: A Windows PC-based system for small-angle scattering data analysis. *Journal of Applied Crystallography* **36**, 1277–1282. ISSN: 00218898 (2003) (cited on pages 47, 123).

-
77. Svergun, D. I., Petoukhov, M. V. & Koch, M. H. Determination of domain structure of proteins from x-ray solution scattering. *Biophysical Journal* **80**, 2946–2953. ISSN: 00063495. [http://dx.doi.org/10.1016/S0006-3495\(01\)76260-1](http://dx.doi.org/10.1016/S0006-3495(01)76260-1) (2001) (cited on pages 48, 123).
 78. Bushue, N. & Wan, Y.-J. Y. Retinoid pathway and cancer therapeutics. *Advanced drug delivery reviews* **62**, 1285–1298. ISSN: 0169-409X. <http://www.sciencedirect.com/science/article/pii/S0169409X10001468> (Oct. 2010) (cited on page 50).
 79. Pérez-Mendoza, D. *et al.* A novel c-di-GMP binding domain in glycosyltransferase BgsA is responsible for the synthesis of a mixed-linkage β -glucan. *Scientific Reports* **7**, 1–11. ISSN: 20452322 (2017) (cited on page 51).
 80. Mercatelli, D., Scalambra, L., Triboli, L., Ray, F. & Giorgi, F. M. Gene regulatory network inference resources: A practical overview. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* **1863**, 194430. ISSN: 18764320. <https://doi.org/10.1016/j.bbagr.2019.194430> (2020) (cited on page 57).
 81. Talbert, P. B., Meers, M. P. & Henikoff, S. Old cogs, new tricks: the evolution of gene expression in a chromatin context. *Nature Reviews Genetics* **20**, 283–297. ISSN: 1471-0056. <http://dx.doi.org/10.1038/s41576-019-0105-7> (May 2019) (cited on pages 57, 104).
 82. Initiative, T. A. G. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815. ISSN: 0028-0836. <http://www.nature.com/articles/35048692> (Dec. 2000) (cited on page 57).
 83. Kim, D. *et al.* Corecognition of DNA and a methylated histone tail by the MSL3 chromodomain. *Nature Structural and Molecular Biology* **17**, 1027–1029. ISSN: 15459993 (2010) (cited on page 57).
 84. Gates, L. A., Foulds, C. E. & O'Malley, B. W. Histone Marks in the 'Driver's Seat': Functional Roles in Steering the Transcription Cycle. *Trends in Biochemical Sciences* **42**, 977–989. ISSN: 13624326. <http://dx.doi.org/10.1016/j.tibs.2017.10.004> (2017) (cited on pages 57, 78).
 85. Yang, X.-J. in *Chromatin Signaling and Diseases* (eds Binda, O. & Fernandez-Zapico, M. E.) 3–23 (Academic Press, Boston, 2016). ISBN: 978-0-12-802389-1. <http://www.sciencedirect.com/science/article/pii/B9780128023891000010> (cited on page 57).
 86. Harp, J. M., Hanson, B. L., Timm, D. E. & Bunick, G. J. Asymmetries in the nucleosome core particle at 2.5 angstrom resolution. *Acta Crystallographica Section D: Biological Crystallography* **56**, 1513–1534 (2000) (cited on pages 57, 58).
 87. Kasinsky, H. E., Lewis, J. D., Dacks, J. B. & Ausió, J. Origin of H1 linker histones. *FASEB Journal* **15**, 34–42. ISSN: 08926638 (2001) (cited on page 58).
 88. Alva, V., Ammelburg, M., Söding, J. & Lupas, A. N. On the origin of the histone fold. *BMC Structural Biology* **7**, 1–10. ISSN: 14726807 (2007) (cited on page 58).
 89. Kale, S., Goncarenco, A., Markov, Y., Landsman, D. & Panchenko, A. R. Molecular recognition of nucleosomes by binding partners. *Current Opinion in Structural Biology* **56**, 164–170. ISSN: 1879033X. <https://doi.org/10.1016/j.sbi.2019.03.010> (2019) (cited on page 58).
 90. Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293**, 1074–1080. ISSN: 00368075 (2001) (cited on page 58).

91. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41–45. ISSN: 0028-0836. <http://www.nature.com/articles/47412> (Jan. 2000) (cited on page 58).
92. Rothbart, S. B., Krajewski, K., Strahl, B. D. & Fuchs, S. M. *Peptide microarrays to interrogate the "histone code"* 1st edition, 107–135. ISBN: 9780123919403. <http://dx.doi.org/10.1016/B978-0-12-391940-3.00006-8> (Elsevier Inc., 2012) (cited on pages 58, 124, 125).
93. Haghbandish, N. & Côté, J. in *Chromatin Signaling and Diseases* 55–74 (2016). ISBN: 9780128026090 (cited on page 58).
94. Eissenberg, J. C. in *Chromatin Signaling and Diseases* 114–124 (Elsevier Inc., 2016). ISBN: 9780128026090. <http://dx.doi.org/10.1016/B978-0-12-802389-1.00006-X> (cited on pages 58–60, 63, 93).
95. Beaver, J. E. & Waters, M. L. Molecular Recognition of Lys and Arg Methylation. *ACS Chemical Biology* **11**, 643–653. ISSN: 1554-8929. <https://pubs.acs.org/doi/10.1021/acscchembio.5b00996> (Mar. 2016) (cited on pages 58, 59, 78).
96. Wang, T. *et al.* Histone variants: critical determinants in tumour heterogeneity. *Frontiers of Medicine* **13**, 289–297. ISSN: 20950225 (2019) (cited on pages 58, 104).
97. Farrona, S., Hurtado, L. & Reyes, J. C. A Nucleosome Interaction Module Is Required for Normal Function of Arabidopsis thaliana BRAHMA. *Journal of Molecular Biology* **373**, 240–250. ISSN: 00222836 (2007) (cited on page 58).
98. Côté, J. & Richard, S. Tudor domains bind symmetrical dimethylated arginines. *Journal of Biological Chemistry* **280**, 28476–28483. ISSN: 00219258 (2005) (cited on page 58).
99. Zhou, Y. *et al.* Clinicopathological significance of ALDH1A1 in lung, colorectal, and breast cancers: a meta-analysis. *Biomarkers in medicine* **9**, 777–90. ISSN: 1752-0371. <http://www.ncbi.nlm.nih.gov/pubmed/26230297> (Aug. 2015) (cited on pages 59–61, 64, 68, 69, 71, 72, 80, 88, 89, 93, 103, 106).
100. Botuyan, M. V. & Mer, G. in *Chromatin Signaling and Diseases* 149–165 (Elsevier Inc., 2016). ISBN: 9780128026090. <http://dx.doi.org/10.1016/B978-0-12-802389-1.00008-3> (cited on pages 59, 64, 68, 93, 106).
101. Maurer-Stroh, S. *et al.* The Tudor domain ‘Royal Family’: Tudor, plant Agenet, Chromo, PWWP and MBT domains. *Trends in Biochemical Sciences* **28**, 69–74. ISSN: 09680004. <https://linkinghub.elsevier.com/retrieve/pii/S0968000403000045> (Feb. 2003) (cited on pages 59–61, 65, 68, 69, 80, 90).
102. Milosevich, N., Warmerdam, Z. & Hof, F. Structural aspects of small-molecule inhibition of methyllysine reader proteins. *Future Medicinal Chemistry* **8**, 1681–1702. ISSN: 17568927 (2016) (cited on pages 59, 71).
103. Sbardella, G. in *Top Med Chem* July, 339–399 (2019). http://link.springer.com/10.1007/7355%7B%5C_%7D2019%7B%5C_%7D78 (cited on page 59).
104. Kaur, G., Iyer, L. M., Subramanian, S. & Aravind, L. Evolutionary convergence and divergence in archaeal chromosomal proteins and Chromo-like domains from bacteria and eukaryotes. *Scientific Reports* **8**, 1–10. ISSN: 20452322. <http://dx.doi.org/10.1038/s41598-018-24467-z> (2018) (cited on pages 59, 60).
105. Eissenberg, J. C. Structural biology of the chromodomain: Form and function. *Gene* **496**, 69–78. ISSN: 03781119. <http://dx.doi.org/10.1016/j.gene.2012.01.003> (2012) (cited on page 60).

106. Nielsen, P. R. *et al.* Structure of the chromo barrel domain from the MOF acetyltransferase. *Journal of Biological Chemistry* **280**, 32326–32331. ISSN: 00219258 (2005) (cited on pages 60, 61, 86, 88, 105, 106).
107. Youkharibache, P. *et al.* The Small β -Barrel Domain: A Survey-Based Structural Analysis. *Structure* **27**, 6–26. ISSN: 18784186. <https://doi.org/10.1016/j.str.2018.09.012> (2019) (cited on page 60).
108. Jacobs, S. A. & Khorasanizadeh, S. Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. *Science* **295**, 2080–2083. ISSN: 00368075 (2002) (cited on pages 60, 61).
109. Zhang, Q. Q., Tian, G. M. & Jin, R. C. The occurrence, maintenance, and proliferation of antibiotic resistance genes (ARGs) in the environment: influencing factors, mechanisms, and elimination strategies. *Applied Microbiology and Biotechnology* **102**, 8261–8274. ISSN: 14320614 (2018) (cited on pages 61, 65, 86, 88, 106).
110. Xu, C., Cui, G., Botuyan, M. V. & Mer, G. Structural Basis for the Recognition of Methylated Histone H3K36 by the Eaf3 Subunit of Histone Deacetylase Complex Rpd3S. *Structure* **16**, 1740–1750. ISSN: 09692126. <http://dx.doi.org/10.1016/j.str.2008.08.008> (2008) (cited on page 62).
111. Shimojo, H. *et al.* Novel Structural and Functional Mode of a Knot Essential for RNA Binding Activity of the Esa1 Presumed Chromodomain. *Journal of Molecular Biology* **378**, 987–1001. ISSN: 00222836 (2008) (cited on page 61).
112. Assland, R. & Stewart, F. The chromo shadow domain, a second chromo domain in heterochromatin-binding protein 1, HP1. *Nucleic Acids Research* **23**, 3168–3173. ISSN: 03051048 (1995) (cited on pages 63, 93).
113. Cowieson, N. P., Partridge, J. F., Allshire, R. C. & McLaughlin, P. J. Dimerisation of a chromo shadow domain and distinctions from the chromodomain as revealed by structural analysis. *Current Biology* **10**, 517–525. ISSN: 09609822 (2000) (cited on page 63).
114. Lu, R. & Wang, G. G. Tudor: A versatile family of histone methylation 'readers'. *Trends in Biochemical Sciences* **38**, 546–555. ISSN: 09680004. <http://dx.doi.org/10.1016/j.tibs.2013.08.002> (2013) (cited on pages 64, 78, 90).
115. Musselman, C. A. *et al.* Molecular basis for H3K36me3 recognition by the Tudor domain of PHF1. *Nature Structural and Molecular Biology* **19**, 1266–1272. ISSN: 15459993. <http://dx.doi.org/10.1038/nsmb.2435> (2012) (cited on pages 65, 101).
116. Adams-Cioaba, M. A. *et al.* Crystal structures of the Tudor domains of human PHF20 reveal novel structural variations on the Royal Family of proteins. *FEBS Letters* **586**, 859–865. ISSN: 00145793. <http://dx.doi.org/10.1016/j.febslet.2012.02.012> (2012) (cited on pages 65, 78, 86, 88, 103, 106).
117. Tang, J. *et al.* Acetylation limits 53BP1 association with damaged chromatin to promote homologous recombination. *Nature Structural and Molecular Biology* **20**, 317–325. ISSN: 15459993. arXiv: arXiv:1507.02142v2. <http://dx.doi.org/10.1038/nsmb.2499> (2013) (cited on pages 66, 78, 94, 101).
118. Zhao, S., Yue, Y., Li, Y. & Li, H. Identification and characterization of 'readers' for novel histone modifications. *Current Opinion in Chemical Biology* **51**, 57–65. ISSN: 13675931. <https://doi.org/10.1016/j.cbpa.2019.04.001> (Aug. 2019) (cited on page 65).
119. Adams-Cioaba, M. A. *et al.* Structural Studies of the Tandem Tudor Domains of Fragile X Mental Retardation Related Proteins FXR1 and FXR2. *PLoS ONE* **5**. ISSN: 19326203 (2010) (cited on page 65).

120. Adinolfi, S. *et al.* The N-Terminus of the Fragile X Mental Retardation Protein Contains a Novel Domain Involved in Dimerization and RNA Binding. *Biochemistry* **42**, 10437–10444. ISSN: 0006-2960. <https://pubs.acs.org/doi/10.1021/bi034909g> (Sept. 2003) (cited on page 65).
121. Lee, J., Thompson, J. R., Botuyan, M. V. & Mer, G. Distinct binding modes specify the recognition of methylated histones H3K4 and H4K20 by JMJD2A-tudor. *Nature Structural and Molecular Biology* **15**, 109–111. ISSN: 15459993 (2008) (cited on pages 67, 101).
122. Liu, S., Jia, J., Gao, Y., Zhang, B. & Han, Y. The AtTudor2, a protein with SN-Tudor domains, is involved in control of seed germination in Arabidopsis. *Planta* **232**, 197–207. ISSN: 00320935 (2010) (cited on pages 68, 88, 93, 106).
123. Gan, B., Chen, S., Liu, H., Min, J. & Liu, K. Structure and function of eTudor domain containing TDRD proteins. *Critical Reviews in Biochemistry and Molecular Biology* **54**, 119–132. ISSN: 15497798. <https://doi.org/10.1080/10409238.2019.1603199> (2019) (cited on pages 68, 103, 106).
124. Bonasio, R., Lecona, E. & Reinberg, D. MBT domain proteins in development and disease. *Seminars in Cell and Developmental Biology* **21**, 221–230. ISSN: 10849521. <http://dx.doi.org/10.1016/j.semcdb.2009.09.010> (2010) (cited on pages 69, 105).
125. Wu, H. *et al.* Structural and Histone Binding Ability Characterizations of Human PWWP Domains. *PLoS ONE* **6** (ed Kursula, P.) e18919. ISSN: 1932-6203. <http://dx.plos.org/10.1371/journal.pone.0018919> (June 2011) (cited on pages 69, 75, 80, 91, 103, 105, 106).
126. Rondelet, G., Dal Maso, T., Willems, L. & Wouters, J. Structural basis for recognition of histone H3K36me3 nucleosome by human de novo DNA methyltransferases 3A and 3B. *Journal of Structural Biology* **194**, 357–367. ISSN: 10958657 (2016) (cited on page 70).
127. Vezzoli, A. *et al.* Molecular basis of histone H3K36me3 recognition by the PWWP domain of Brpf1. *Nature Structural and Molecular Biology* **17**, 617–619. ISSN: 15459993 (2010) (cited on page 69).
128. Pascual, J., Martinez-Yamout, M., Dyson, H. J. & Wright, P. E. Structure of the PHD zinc finger from human Williams-Beuren syndrome transcription factor. *Journal of Molecular Biology* **304**, 723–729. ISSN: 00222836 (2000) (cited on page 71).
129. Li, H. *et al.* Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature* **442**, 91–95. ISSN: 14764687 (2006) (cited on page 71).
130. Morrison, E. A. & Musselman, C. A. in *Chromatin Signaling and Diseases* 127–147 (Elsevier Inc., 2016). ISBN: 9780128026090. <http://dx.doi.org/10.1016/B978-0-12-802389-1.00007-1> (cited on page 71).
131. Lan, F. *et al.* Recognition of unmethylated histone H3 lysine 4 links BHC80 to LSD1-mediated gene repression. *Nature* **448**, 718–722. ISSN: 0028-0836. <http://www.nature.com/articles/nature06034> (Aug. 2007) (cited on page 71).
132. Peña, P. V. *et al.* Molecular mechanism of histone H3K4me3 recognition by plant homeodomain of ING2. *Nature* **442**, 100–103. ISSN: 14764687 (2006) (cited on page 72).
133. Suganuma, T., Pattenden, S. G. & Workman, J. L. Diverse functions of WD40 repeat proteins in histone recognition. *Genes and Development* **22**, 1265–1268. ISSN: 08909369 (2008) (cited on pages 71, 72).
134. Xu, C. & Min, J. Structure and function of WD40 domain proteins. *Protein and Cell* **2**, 202–214. ISSN: 16748018 (2011) (cited on pages 71, 72).

135. Jain, B. P. & Pandey, S. WD40 Repeat Proteins: Signalling Scaffold with Diverse Functions. *Protein Journal* **37**, 391–406. ISSN: 15734943. <http://dx.doi.org/10.1007/s10930-018-9785-7> (2018) (cited on pages 71, 72).
136. Paoli, M. Protein folds propelled by diversity. *Progress in Biophysics and Molecular Biology* **76**, 103–130. ISSN: 00796107. <https://linkinghub.elsevier.com/retrieve/pii/S0079610701000074> (2001) (cited on page 71).
137. Van Nocker, S. & Ludwig, P. The WD-repeat protein superfamily in Arabidopsis: Conservation and divergence in structure and function. *BMC Genomics* **4**, 1–11. ISSN: 14712164 (2003) (cited on page 72).
138. Couture, J. F., Collazo, E. & Trievel, R. C. Molecular recognition of histone H3 by the WD40 protein WDR5. *Nature Structural and Molecular Biology* **13**, 698–703. ISSN: 15459993 (2006) (cited on pages 72, 73).
139. Holm, L. & Laakso, L. M. Dali server update. *Nucleic acids research* **44**, W351–W355. ISSN: 13624962 (2016) (cited on page 78).
140. Huang, Y., Fang, J., Bedford, M. T., Zhang, Y. & Xu, R.-m. Recognition of Histone H3 Lysine-4 Methylation by the Double Tudor Domain of JMJD2A. *Science* **312**, 748–751. ISSN: 0036-8075 (2006) (cited on pages 78, 80, 101, 103, 106).
141. Sonnhammer, E. L., Eddy, S. R. & Durbin, R. Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins: Structure, Function and Genetics* **28**, 405–420. ISSN: 08873585 (1997) (cited on page 78).
142. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Research* **47**, D427–D432. ISSN: 13624962 (2019) (cited on page 78).
143. Mitchell, A. L. *et al.* InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research* **47**, D351–D360. ISSN: 0305-1048. <https://academic.oup.com/nar/article/47/D1/D351/5162469> (Jan. 2019) (cited on page 78).
144. Potter, S. C. *et al.* HMMER web server: 2018 update. *Nucleic Acids Research* **46**, W200–W204. ISSN: 13624962 (2018) (cited on page 78).
145. Nielsen, P. R., Callaghan, J., Murzin, A. G., Murzina, N. V. & Laue, E. D. Expression, Purification, and Biophysical Studies of Chromodomain Proteins. *Methods in Enzymology* **376**, 148–170. ISSN: 00766879 (2004) (cited on pages 80, 91).
146. Crooks, G. E. WebLogo: A Sequence Logo Generator. *Genome Research* **14**, 1188–1190. ISSN: 1088-9051. <http://www.genome.org/cgi/doi/10.1101/gr.849004> (May 2004) (cited on page 84).
147. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**. ISSN: 17444292 (2011) (cited on page 84).
148. Ashkenazy, H. *et al.* ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic acids research* **44**, W344–W350. ISSN: 13624962 (2016) (cited on page 84).
149. Celniker, G. *et al.* ConSurf: Using evolutionary data to raise testable hypotheses about protein function. *Israel Journal of Chemistry* **53**, 199–206. ISSN: 00212148 (2013) (cited on page 84).

150. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: Calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Research* **38**, W529–W533. ISSN: 03051048. <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq399> (July 2010) (cited on page 84).
151. Landau, M. *et al.* ConSurf 2005: The projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Research* **33**, 299–302. ISSN: 03051048 (2005) (cited on page 84).
152. Glaser, F. *et al.* ConSurf: Identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* **19**, 163–164. ISSN: 13674803 (2003) (cited on page 84).
153. Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S. & Brunak, S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *PROTEOMICS* **4**, 1633–1649. ISSN: 1615-9853. <http://doi.wiley.com/10.1002/pmic.200300771> (June 2004) (cited on page 85).
154. Blom, N., Gammeltoft, S. & Brunak, S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *Journal of Molecular Biology* **294**, 1351–1362. ISSN: 00222836. <https://linkinghub.elsevier.com/retrieve/pii/S0022283699933107> (Dec. 1999) (cited on page 85).
155. Cui, G. *et al.* PHF20 is an effector protein of p53 double lysine methylation that stabilizes and activates p53. *Nature Structural and Molecular Biology* **19**, 916–924. ISSN: 15459993. <http://dx.doi.org/10.1038/nsmb.2353> (2012) (cited on pages 88, 90).
156. Sprangers, R., Groves, M. R., Sinning, I. & Sattler, M. High-resolution X-ray and NMR structures of the SMN Tudor domain: Conformational variation in the binding site for symmetrically dimethylated arginine residues. *Journal of Molecular Biology* **327**, 507–520. ISSN: 00222836 (2003) (cited on pages 88, 106).
157. Tripsianes, K. *et al.* Structural basis for dimethylarginine recognition by the Tudor domains of human SMN and SPF30 proteins. *Nature Structural and Molecular Biology* **18**, 1414–1420. ISSN: 15459993 (2011) (cited on pages 88, 106).
158. Guo, Y. *et al.* Methylation-state-specific recognition of histones by the MBT repeat protein L3MBTL2. *Nucleic Acids Research* **37**, 2204–2210. ISSN: 03051048 (2009) (cited on pages 89, 106).
159. Liu, J. *et al.* Structural plasticity of the TDRD3 Tudor domain probed by a fragment screening hit. *FEBS Journal* **285**, 2091–2103. ISSN: 17424658 (2018) (cited on pages 89, 103, 106).
160. Liu, K. *et al.* Crystal Structure of TDRD3 and Methyl-Arginine Binding Characterization of TDRD3, SMN and SPF30. *PLoS ONE* **7** (ed Gay, N.) e30375. ISSN: 1932-6203. <https://dx.plos.org/10.1371/journal.pone.0030375> (Feb. 2012) (cited on page 89).
161. Persson, H. *et al.* CDR-H3 diversity is not required for antigen recognition by synthetic antibodies. *Journal of Molecular Biology* **425**, 803–811. ISSN: 10898638. <http://dx.doi.org/10.1016/j.jmb.2012.11.037> (2013) (cited on page 89).
162. Sikorsky, T. *et al.* Recognition of asymmetrically dimethylated arginine by TDRD3. *Nucleic Acids Research* **40**, 11748–11755. ISSN: 03051048 (2012) (cited on pages 89, 90).
163. Krissinel, E. & Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D: Biological Crystallography* **60**, 2256–2268. ISSN: 09074449 (2004) (cited on page 90).

-
164. Winn, M. D. *et al.* Overview of the <i>CCP</i> 4 suite and current developments. *Acta Crystallographica Section D Biological Crystallography* **67**, 235–242. ISSN: 0907-4449. <http://scripts.iucr.org/cgi-bin/paper?S0907444910045749> (2011) (cited on pages 90, 122).
 165. Freire Rios, A. *Structural basis for specific gene regulation by Auxin Response Factors* PhD thesis (Wageningen University and Research, 2016), 190. ISBN: 9789462579538. <https://library.wur.nl/WebQuery/wurpubs/510787> (cited on page 93).
 166. Wilson, M. D. *et al.* The structural basis of modified nucleosome recognition by 53BP1. *Nature* **536**, 100–103. ISSN: 14764687 (2016) (cited on page 94).
 167. Botuyan, M. V. *et al.* Structural Basis for the Methylation State-Specific Recognition of Histone H4-K20 by 53BP1 and Crb2 in DNA Repair. *Cell* **127**, 1361–1373. ISSN: 00928674 (2006) (cited on pages 94, 103, 106).
 168. Fradet-Turcotte, A. *et al.* 53BP1 is a reader of the DNA-damage-induced H2A Lys 15 ubiquitin mark. *Nature* **499**, 50–54. ISSN: 00280836 (2013) (cited on page 94).
 169. Katz, C. *et al.* Studying protein-protein interactions using peptide arrays. *Chemical Society Reviews* **40**, 2131–2145. ISSN: 03060012 (2011) (cited on page 95).
 170. Louche, A., Salcedo, S. P. & Bigot, S. in *Bacterial Protein Secretion Systems* 247–255 (2017). ISBN: 978-1-4939-7031-5. <http://link.springer.com/10.1007/978-1-4939-7033-9> (cited on page 96).
 171. Tripathi, L. P., Esaki, T., Itoh, M. N., Chen, Y.-A. & Mizuguchi, K. in *Encyclopedia of Bioinformatics and Computational Biology* 194–199 (Elsevier, 2019). ISBN: 9780128114322. <http://dx.doi.org/10.1016/B978-0-12-809633-8.20096-4> (cited on page 96).
 172. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and Biology Insights* **14**, 117793221989905. ISSN: 1177-9322. <http://journals.sagepub.com/doi/10.1177/1177932219899051> (Jan. 2020) (cited on page 96).
 173. Leung, K. M. Joining the dots between omics and environmental management. *Integrated Environmental Assessment and Management* **14**, 169–173. ISSN: 15513777. <http://doi.wiley.com/10.1002/ieam.2007> (Mar. 2018) (cited on page 96).
 174. Maroli, A. S. *et al.* Omics in Weed Science: A Perspective from Genomics, Transcriptomics, and Metabolomics Approaches. *Weed Science* **66**, 681–695. ISSN: 1550-2759. https://www.cambridge.org/core/product/identifier/S0043174518000334/type/journal%7B%5C_%7Darticle (Nov. 2018) (cited on page 96).
 175. Narad, P. & Kirthanashri, S. V. in *Omics Approaches, Technologies And Applications* 1–10 (Springer Singapore, Singapore, 2018). ISBN: 9789811329258. http://link.springer.com/10.1007/978-981-13-2925-8%7B%5C_%7D1 (cited on page 96).
 176. Boja, E. S., Kinsinger, C. R., Rodriguez, H. & Srinivas, P. Integration of omics sciences to advance biology and medicine. *Clinical Proteomics* **11**, 1–12. ISSN: 15590275 (2014) (cited on page 96).
 177. Brady, A. E., Chen, Y., Limbird, L. E. & Wang, Q. in *Methods in molecular biology (Clifton, N.J.)* May, 347–355 (2011). ISBN: 978-1-61779-125-3. <http://link.springer.com/10.1007/978-1-61779-126-0> (cited on pages 96, 97).

178. Volkmer, R. & Tapia, V. Exploring Protein-Protein Interactions with Synthetic Peptide Arrays. *Mini-Reviews in Organic Chemistry* **8**, 164–170. ISSN: 1570193X. <http://www.eurekaselect.com/openurl/content.php?genre=article%7B%5C%7Dissn=1570-193X%7B%5C%7Dvolume=8%7B%5C%7Disssue=2%7B%5C%7Dspage=164> (May 2011) (cited on page 96).
179. Mauser, R. & Jeltsch, A. Application of modified histone peptide arrays in chromatin research. *Archives of Biochemistry and Biophysics* **661**, 31–38. ISSN: 10960384. <https://doi.org/10.1016/j.abb.2018.10.019> (2019) (cited on pages 98, 124).
180. Hirschev, M. D. & Zhao, Y. Metabolic regulation by lysine malonylation, succinylation, and glutarylation. *Molecular and Cellular Proteomics* **14**, 2308–2315. ISSN: 15359484 (2015) (cited on pages 100, 105).
181. Chen, Y. *et al.* Lysine propionylation and butyrylation are novel post-translational modifications in histones. *Molecular and Cellular Proteomics* **6**, 812–819. ISSN: 15359476 (2007) (cited on pages 100, 105).
182. Ballaré, C. *et al.* Phf19 links methylated Lys36 of histone H3 to regulation of Polycomb activity. *Nature Structural & Molecular Biology* **19**, 1257–1265. ISSN: 1545-9993. <http://www.nature.com/articles/nsmb.2434> (Dec. 2012) (cited on page 101).
183. Bian, C. *et al.* Sgf29 binds histone H3K4me2/3 and is required for SAGA complex recruitment and histone H3 acetylation. *The EMBO Journal* **30**, 2829–2842. ISSN: 02614189. <http://dx.doi.org/10.1038/emboj.2011.193> (July 2011) (cited on page 101).
184. Yang, N. *et al.* Distinct mode of methylated lysine-4 of histone H3 recognition by tandem tudor-like domains of Spindlin1. *Proceedings of the National Academy of Sciences* **109**, 17954–17959. ISSN: 0027-8424. <http://www.pnas.org/cgi/doi/10.1073/pnas.1208517109> (Oct. 2012) (cited on page 101).
185. Arita, K. *et al.* Recognition of modification status on a histone H3 tail by linked histone reader modules of the epigenetic regulator UHRF1. *Proceedings of the National Academy of Sciences* **109**, 12950–12955. ISSN: 0027-8424. <http://www.pnas.org/cgi/doi/10.1073/pnas.1203701109> (Aug. 2012) (cited on page 101).
186. Law, J. A. *et al.* Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. *Nature* **498**, 385–389. ISSN: 00280836. arXiv: NIHMS150003. <http://dx.doi.org/10.1038/nature12178> (2013) (cited on page 101).
187. Kungulovski, G., Kycia, I., Mauser, R. & Jeltsch, A. Specificity Analysis of Histone Modification-Specific Antibodies or Reading Domains on Histone Peptide Arrays. *Methods in molecular biology (Clifton, N.J.)* **1348**, 275–84. ISSN: 1940-6029. <http://www.springer.com/gb/book/9781493929986> (2015) (cited on pages 103, 106).
188. Copeland, R. A. *Enzymes: a practical introduction to structure, mechanism, and data analysis* 2nd editio, 397. ISBN: 0471359297 (Wiley-VCH, Inc, 2000) (cited on page 103).
189. Biterge, B. & Schneider, R. Histone variants: key players of chromatin. *Cell and Tissue Research* **356**, 457–466. ISSN: 0302-766X. <http://link.springer.com/10.1007/s00441-014-1862-4> (June 2014) (cited on page 104).
190. Jiang, D. & Berger, F. Histone variants in plant transcriptional regulation. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* **1860**, 123–130. ISSN: 18764320. <http://dx.doi.org/10.1016/j.bbagr.2016.07.002> (2017) (cited on page 104).
191. Buschbeck, M. *et al.* The histone variant macroH2A is an epigenetic regulator of key developmental genes. *Nature Structural & Molecular Biology* **16**, 1074–1079. ISSN: 1545-9993. <http://www.nature.com/articles/nsmb.1665> (Oct. 2009) (cited on page 104).

192. Yelagandula, R. *et al.* The Histone Variant H2A.W Defines Heterochromatin and Promotes Chromatin Condensation in Arabidopsis. *Cell* **158**, 98–109. ISSN: 00928674. <https://linkinghub.elsevier.com/retrieve/pii/S0092867414007272> (July 2014) (cited on page 104).
193. Talbert, P. B. & Henikoff, S. Histone variants on the move: substrates for chromatin dynamics. *Nature Reviews Molecular Cell Biology* **18**, 115–126. ISSN: 1471-0072. <http://www.nature.com/articles/nrm.2016.148> (Feb. 2017) (cited on page 104).
194. Hyun, K., Jeon, J., Park, K. & Kim, J. Writing, erasing and reading histone lysine methylations. *Experimental and Molecular Medicine* **49**, e324–22. ISSN: 20926413. <http://dx.doi.org/10.1038/emm.2017.11> (2017) (cited on page 104).
195. Herold, J. M., Ingerman, L. A., Gao, C. & Frye, S. V. Drug discovery toward antagonists of methyl-lysine binding proteins. *Current chemical genomics* **5**, 51–61. ISSN: 1875-3973. <http://www.scopus.com/inward/record.url?eid=2-s2.0-80455131079%7B%5C%7DpartnerID=tZ0tx3y1> (2011) (cited on page 105).
196. Zhao, D., Li, Y., Xiong, X., Chen, Z. & Li, H. YEATS Domain—A Histone Acylation Reader in Health and Disease. *Journal of Molecular Biology* **429**, 1994–2002. ISSN: 10898638. <http://dx.doi.org/10.1016/j.jmb.2017.03.010> (2017) (cited on page 105).
197. Smith, S. G. & Zhou, M. M. The Bromodomain: A New Target in Emerging Epigenetic Medicine. *ACS Chemical Biology* **11**, 598–608. ISSN: 15548937 (2016) (cited on page 105).
198. Simpson, L. W., Good, T. A. & Leach, J. B. Protein folding and assembly in confined environments: Implications for protein aggregation in hydrogels and tissues. *Biotechnology Advances* **42**. ISSN: 07349750 (2020) (cited on pages 105, 106).
199. Komatsu, T. & Urano, Y. Chemical toolbox for 'live' biochemistry to understand enzymatic functions in living systems. *Journal of Infectious Diseases* **220**, 139–149. ISSN: 15376613 (2019) (cited on page 105).
200. Ellis, R. J. Macromolecular crowding: Obvious but underappreciated. *Trends in Biochemical Sciences* **26**, 597–604. ISSN: 09680004 (2001) (cited on page 105).
201. Ostrowska, N., Feig, M. & Trylska, J. Modeling Crowded Environment in Molecular Simulations. *Frontiers in Molecular Biosciences* **6**, 1–6. ISSN: 2296889X (2019) (cited on page 105).
202. Lee, D., Redfern, O. & Orengo, C. Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology* **8**, 995–1005. ISSN: 14710072 (2007) (cited on page 106).
203. Simithy, J., Sidoli, S. & Garcia, B. A. Integrating Proteomics and Targeted Metabolomics to Understand Global Changes in Histone Modifications. *Proteomics* **18**, 1–8. ISSN: 16159861 (2018) (cited on page 106).
204. Kungulovski, G. *et al.* Application of histone modification-specific interaction domains as an alternative to antibodies. *Genome Research* **24**, 1842–1853. ISSN: 15495469 (2014) (cited on pages 106, 107).
205. Jeltsch, A. & Kungulovski, G. Quality of histone modification antibodies undermines chromatin biology research. *F1000Research* **4**, 1–12. ISSN: 1759796X (2015) (cited on page 106).
206. Lanctot, A., Taylor-Teeple, M., Oki, E. A. & Nemhauser, J. L. Specificity in Auxin Responses Is Not Explained by the Promoter Preferences of Activator ARFs. *Plant Physiology* **182**, 1533–1536. ISSN: 0032-0889. <http://www.plantphysiol.org/lookup/doi/10.1104/pp.19.01474> (Apr. 2020) (cited on page 112).

207. Juanhuix, J. *et al.* Developments in optics and performance at BL13-XALOC, the macromolecular crystallography beamline at the Alba Synchrotron. *Journal of Synchrotron Radiation* **21**, 679–689. ISSN: 16005775 (2014) (cited on pages 121, 122).
208. Kabsch, W. *XDS*. *Acta Crystallographica Section D Biological Crystallography* **66**, 125–132. ISSN: 0907-4449. <http://scripts.iucr.org/cgi-bin/paper?S0907444909047337> (2010) (cited on page 121).
209. Vonrhein, C. *et al.* Data processing and analysis with the autoPROC toolbox. *Acta Crystallographica Section D: Biological Crystallography* **67**, 293–302. ISSN: 09074449 (2011) (cited on page 122).
210. McCoy, A. J. *et al.* Phaser crystallographic software. *Journal of applied crystallography* **40**, 658–674. ISSN: 0021-8898 (Aug. 2007) (cited on page 122).
211. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta crystallographica. Section D, Biological crystallography* **66**, 213–221. ISSN: 1399-0047 (Feb. 2010) (cited on page 122).
212. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta crystallographica. Section D, Biological crystallography* **66**, 486–501. ISSN: 1399-0047 (Apr. 2010) (cited on page 122).
213. Williams, C. J. *et al.* MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science* **27**, 293–315. ISSN: 1469896X (2018) (cited on page 122).
214. Franke, D. *et al.* ATSAS 2.8: A comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *Journal of Applied Crystallography* **50**, 1212–1225. ISSN: 16005767 (2017) (cited on page 123).
215. Svergun, D., Barberato, C. & Koch, M. H. CRY SOL - A program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *Journal of Applied Crystallography* **28**, 768–773. ISSN: 00218898 (1995) (cited on page 123).
216. Kozin, M. B. & Svergun, D. I. Automated matching of high- and low-resolution structural models. *Journal of Applied Crystallography* **34**, 33–41. ISSN: 00218898 (2001) (cited on page 123).
217. Panjkovich, A. & Svergun, D. I. Deciphering conformational transitions of proteins by small angle X-ray scattering and normal mode analysis. *Physical Chemistry Chemical Physics* **18**, 5707–5719. ISSN: 14639076 (2016) (cited on page 123).
218. Fuchs, S. M., Krajewski, K., Baker, R. W., Miller, V. L. & Strahl, B. D. Influence of combinatorial histone modifications on antibody and effector protein recognition. *Current Biology* **21**, 53–58. ISSN: 09609822. <http://dx.doi.org/10.1016/j.cub.2010.11.058> (2011) (cited on page 124).
219. Shanle, E. K. *et al.* Histone peptide microarray screen of chromo and Tudor domains defines new histone lysine methylation interactions. *Epigenetics and Chromatin* **10**, 1–11. ISSN: 17568935 (2017) (cited on page 124).
220. Rossi, A. M. & Taylor, C. W. Analysis of protein-ligand interactions by fluorescence polarization. *Nature Protocols* **6**, 365–387. ISSN: 17542189. <http://dx.doi.org/10.1038/nprot.2011.305> (2011) (cited on page 125).
221. Mahmood, T. & Yang, P. C. Western blot: Technique, theory, and trouble shooting. *North American Journal of Medical Sciences* **4**, 429–434. ISSN: 22501541 (2012) (cited on page 126).

-
222. Su, M. Y. *et al.* Structural basis of adaptor-mediated protein degradation by the tail-specific PDZ-protease *Proc. Nature Communications* **8**, 1–13. ISSN: 20411723. <http://dx.doi.org/10.1038/s41467-017-01697-9> (2017).

Publications

1. H. Kato, S.K. Mutte*, H. Suzuki*, I. Crespo*, S. Das*, T. Radoeva*, M. Fontana*, Y. Yoshitake, E. Hainiwa, W. van den Berg, S. Lindhoud, K. Ishizaki, J. Hohlbein, J.W. Borst, R. Boer, R. Nishihama, T. Kohchi, K. Ishizaki, D. Weijers, **Design principles of a minimal auxin response system**, (2020) *Nature Plants*, 6(5), 473–482. <https://doi.org/10.1038/s41477-020-0662-y>. *SKM, HS, IC, SD, TR and MF contributed equally to this work
2. A. Freire-Rios*, K. Tanaka*, I. Crespo, E. van Wijk, Y. Sizensova, V. Levitsky, S. Lindhoud, M. Fontana, J. Hohlbein, D.R. Boer, V. Mironova, D. Weijers, **Architecture of DNA elements mediating ARF transcription factor binding and auxin-responsive gene expression in Arabidopsis**, (2020). *Proceedings of the National Academy of Sciences* Sep 2020, 117 (39) 24557-24566; <https://doi.org/10.1073/pnas.2009554117>. *AFR and KT contributed equally to this work.
3. **Refining the molecular calliper model of ARF gene selection from evolutionary contributions**. In preparation. This paper consists of the findings presented in chapters 2 and 3.
4. **Steward Domain: the connection between epigenetics and auxin signalling**. In preparation. This paper will consist of the findings presented in chapters 5 and 6.
5. **The Royal Family: a complex family of histone PTM readers**. In preparation. This paper will consist in the review presented in Chapter 4.

Deposited structures (PDBs)

1. MpARF2-DBD:21ds, $C2$, 2.96 Å, PDBid: 6SDG
2. AtARF1-DBD:21ds, $P2_1$, 1.65 Å, PDBid: 6YCQ
3. MpARF2-DBD:21ds, $I2_12_12_1$, 2.32 Å, PDBid: *Pending assignment*
4. MpARF2-DBD:ER7, $I2_12_12_1$, 2.56 Å, PDBid: *Pending assignment*