DISS. No. ETH 21877

NEW INSIGHTS INTO THE DETERMINANTS OF PROTEIN-RNA INTERACTIONS:

A structural, thermodynamic and kinetic investigation of the hnRNP C and UNR proteins

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

ZUZANA CIENIKOVÁ

Ing. dipl. INSA Lyon
born November 06, 1982
citizen of Slovakia

accepted on the recommendation of

Prof. Dr. Frédéric ALLAIN
Prof. Dr. Nenad BAN
Prof. Dr. Sebastian HILLER

2014

Summary

In the central dogma of molecular biology, RNA is a messenger molecule carrying the genetic information from DNA to proteins. However, this simplified view leaves many functions of RNA out of the picture. Today, it is clear that the ribonucleic acids steer heavily the fate of a cell, be it as extensively regulated messenger molecules or active core components of the most important molecular machines. To gain new insights into the determinants of RNA metabolism, this work investigated the RNA interactions of two human proteins, UNR and hnRNP C. UNR is a cytoplasmic protein acting as translation regulator. Binding to mRNA is achieved by a combinatorial interplay of its five *cold-shock domains* [CSD]. Using NMR, we show that the first domain specifically binds the known consensus motifs of the full-length protein, employing the canonical interaction surface of CSDs. HnRNP C is a very abundant cellular protein involved at all stages of mRNA processing. The protein contains one *RNA recognition motif* [RRM] known to bind uridine-rich sequences. To understand the molecular basis of this recognition, we solved the solution structures of this RRM in complex with poly-uridine oligomers of two different lengths. The structures reveal that the RRM is able to bind up to five uridines, with a gradient of discriminating power increasing in the 5' to 3' direction. The core of the specific recognition lies outside of the RRM, in the N- and C-terminal extensions of the domain. The sequence degeneracy of the binding consensus shapes this domain's propensity to multivalent binding on poly-uridine tracts, leading to short uridine tract saturation pattern remarkably consistent with the *in-vivo* RNA crosslink distribution obtained with the full-length hnRNP C. This result re-establishes the RRM as the key RNA-binding domain of hnRNP C tetramer. We show furthermore with a combination of *in-vivo,* biophysical and structural modelling approaches that extended uridine tracts can accommodate two RRMs side-by-side. This result suggests that two hnRNP C RRMs coordinated through the tetramer positively cooperate to bind long tracts in tandem. This behavior appears to be a general feature of hnRNP C RNA recognition, since the transcriptome-wide hnRNP C crosslinking pattern fits well with a model of dual RRM binding on long uridine tracts, with a 14-fold affinity enhancement compared to the affinity of single RRM binding event.

Résumé

D'après le dogme central de la biologie moléculaire, l'ARN est le porteur intermédiaire de l'information génétique entre l'ADN et les protéines. Cependant cette vue simplifiée ne tient pas compte de nombreuses autres fonctions de l'ARN. De nos jours, il est communément admis que les acides ribonucléiques influencent le devenir de la cellule; que ce soit en tant que messager ou en tant que composant principal d'importantes machines moléculaires. Afin d'avoir une meilleure compréhension du métabolisme de l'ARN, ces travaux se sont portés sur l'interaction entre l'ARN et les protéines humaines UNR et hnRNP C.

UNR est une protéine cytoplasmique régulatrice de la traduction. Sa liaison à l'ARNm se fait par l'intermédiaire de ses cinq domaines de choc au froid. Grâce à la RMN, nous avons démontré que le premier domaine se lie spécifiquement aux séquences consensus connues pour la protéine entière, en utilisant la surface d'interaction canonique des domaines de choc au froid.

HnRNP C est une protéine très abondante, impliquée dans toutes les étapes de maturation de l'ARNm. Cette protéine contient un domaine de reconnaissance de l'ARN [RRM] connu pour se lier à des séquences riches en uridines. Pour comprendre les bases moléculaires de cette reconnaissance, nous avons résolu les structures en solution de ce RRM en complexe avec des oligomères poly-uridine de deux longueurs différentes. Ces structures mettent en évidence l'habilité du RRM à lier jusqu'à cinq uridines, avec un pouvoir discriminant croissant de 5' vers 3'. La spécificité est apportée par les extensions N- et C-terminales et non le RRM. Le caractère dégénéré du consensus explique la propension du domaine à la liaison multi-registre sur les chaînes de poly-uridines, conduisant à une saturation des sites poly(U) courts remarquablement en accord avec la distribution *in-vivo* des « UV-crosslinks » entre l'ARN et la protéine hnRNP C entière. Ce résultat rétablit le RRM comme le domaine clé de la reconnaissance de l'ARN par la forme tétramèrique de hnRNP C. De plus, en combinant des approches *in-vivo*, biophysiques et de modélisation structurale, nous avons démontré que des sites poly-uridine étendus peuvent accueillir deux RRMs accolés. Nos résultats suggèrent en plus une coopérativité positive entre deux RRMs de hnRNP C coordonnés par le tétramère lors de la liaison aux longues U-chaînes. Ce comportement semble être une caractéristique générale de la liaison de hnRNP C à l'ARN, puisque le résultat de « crosslinking » pan-transcriptomique de hnRNP C corrèle bien avec le modèle de liaison coopérative sur de longs sites uridine, avec une multiplication par 14 de l'affinité comparée à la liaison d'un seul RRM.