UNIVERSITÀ
DEGLI STUDI
FIRENZE

DOTTORATO DI RICERCA
INTERNATIONAL DOCTORATE IN STRUCTURAL BIOLOGY

CICLO XXVIII

COORDINATORE Prof. Claudio Luchinat

Characterization of heterogeneous viral proteins by NMR methods: Human adenovirus E1A and human papillomavirus E7 proteins

Settore Scientifico Disciplinare CHIM/03

Dottorando
Dott. Marcela Nogueira

Tutore
Prof. Isabella C. Felli

Co-Tutore
Prof. Roberta Pierattelli

Coordinatore
Prof. Claudio Luchinat

November 2012-2016

Universiteit Utrecht

CERM Firenze

Acknowledgements

During the development of this work I had the opportunity to learn so many different things. This period of my life was essential to grow as scientist and as a person. I would like also to thank here all the people who supported me in this process and all those who contributed to this thesis.

First of all I would like to express my gratitude to the Brazilian government and to Dr. Carlos Medicis Morel for the opportunity given to me to study and get knowledge within the frame of the Science Without Borders program. I hope to be able to repay this chance applying in my country all the lived experience.

I would like to acknowledge my supervisors Isabella C. Felli and Roberta Pierattelli for the opportunity to work at CERM. Thank you for the support, guidance and teaching. Thanks also for all the time you have invested with discussions and suggestions during my PhD. It was very important to make me feel more and more confident with the work.

I thank Dr. Jochen Junker for encouraging me always to move on. Thank you for your professional and personal support during all these years we have worked together. I will always be grateful to you for your support.

I would like to acknowledge the scientific board of teachers of CERM for the interesting scientific discussions in the seminar group and lessons. I thank also Dr. Francesca Cantini for all discussions and help on the bioinformatic part.

I would like to thank all the technicians. Specially, Fabio and Massimo who were so helpful and for all discussions and patience you had. I thank also Rebecca for organizing the NMR schedule time and making changes even at the last moment to leave me machine time. I thank to Leonardo for organizing the life of the wet lab and for the help. I learned precious things with your advice.

I would like also to acknowledge my closest friends from the lab who helped me a lot during the PhD. Eduardo, who was there at the beginning introduced me to the work in the wet-lab. It was a pleasure to work with him and share so many jokes and laughs; Tomás, who taught me many things about the spectrometer and about bioinformatics programs. Thank you for all the professional and personal support. I

will never forget your patience and kindness; and Linda, who was always so helpful and open for discussions. Thanks for all the time and effort you gave to me.

I thank also my fellow in the lab, Francesca Castiglia, with whom I had the opportunity to teach what I learned and had a chance to improve with her my Italian language. Many Thanks to you Tommaso, Azzurra, Talita, Witold, Stefano, Angelo, Alessandro and Daniela for all contributions in the lab and friendship. I thank also Roberto for all help, for the kindness and care he had to me. Thanks to all of you for the funny moments!

This acknowledgement would be incomplete if I would not mention my beloved friends Maria Cristina and Paulo Baía for all support and encouragement they always give me. I also thank Sarah, Ludo and Julien, who were there for me with their kindness, providing me the moral support. Thanks for all great moments we have spent together.

Finally and most importantly, I owe to express my heartfelt gratitude to my family, in particular to my parents, who supported me in doing what I wanted to do, even in moving so far away from them for enrolling in this PhD program. I acknowledge also my parents-in-law for giving me all the support, care and affection. Most especially I thank my husband, Benoît who has been by my side, with his kindness and unconditional love at each and every step of my journey, participating in good and bad moments. Without his support this would not be possible. Thanks for everything. This thesis is dedicated to him.

Abstract

Intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) are characterized by high disorder, flexibility and lack a stable three-dimensional structure. These properties confer to IDPs advantages for molecular recognition and interactions. In fact, IDPs and IDRs often act as hub proteins, by interacting with different partners through a unique segment. Since IDPs and IDRs are highly flexible and more exposed, they can be readily modified through post-translational modifications, heaving thus, an additional point for modulation of their activities. These particular features offered by IDPs and IDRs are extremely relevant for viruses, which need to overcome the limitation of their small genome that is not large enough to encode all proteins necessary to interact with key proteins from the host cell.

Intrinsic disorder is relatively abundant among cancer-related diseases where IDPs play a number of key roles in oncogenesis. Good examples of oncogenic viral proteins are E7 from human papillomavirus (HPV) and E1A from human adenovirus (HAdV). In this work we present the structural and dynamic characterization, at atomic resolution of both E7 and E1A proteins. This is an important step to better understand the particularities of these proteins, laying the basis to study important interactions.

E7 and E1A proteins act as molecular hubs, interacting with many key partners like the retinoblastoma tumour suppressor (pRb) and p300/CREB-biding protein (CBP), for example. Many of the interactions of both proteins are modulated by PTMs, in particular by their phosphorylation state, which has been shown to increase the affinity with their partners. Although the biological importance of phosphorylation events on E7 and E1A is well known, the structural effect of this PTM is still not clear. In this context, as a first step, we also investigated the structural effect of one of the most important PTMs on HPV-16 E7, the phosphorylation by Casein Kinase II.

Keywords

Intrinsically disordered proteins, IDPs, HPV-16 E7, HAdV E1A, NMR, post-translational modifications, PTMs, phosphorylation, CKII, conserved region.

Table of Contents

Acknowledgements i

Abstract iii

Keywords iv

Table of Contents v

List of Abbreviations vii

Chapter 1. Introduction 1

1.1 E7 protein from human papillomavirus 5

1.2 Human adenovirus E1A proteins 11

1.3 Post-translational modifications 17

1.3.1 PTMs of E1A proteins 18

1.3.2 PTMs of E7 19

1.4 Casein Kinase type II and its requirements to perform phosphorylation 21

1.5 Studying PTMs by NMR techniques 23

Chapter 2. Methods 27

2.1 Bioinformatics studies 27

2.1.1 Prediction of disordered proteins/domains 28

2.2 Recombinant protein expression using *Escherichia coli* systems 29

2.3 The plasmid systems 30

2.3.1 Plasmid of the full-length HPV-16 E7 31

2.3.2 Plasmid of E7CR3 and E1ACR3 domains 32

2.4 Cell transformation 33

2.5 Expression tests and final cultures 34

2.6 Protein purification strategy 37

2.7 Refolding Processes 42

2.8 Structural and dynamic properties of proteins by NMR techniques 44

2.9 Sequence specific assignment and structure determination 46

2.10 Nuclear Overhauser effect 48

2.11 Residual Dipolar Coupling 49
2.11.1 Alignment media 50
2.12 Dynamic studies 53
2.13 Computational approaches to describe the structural and dynamic properties of IDPs 54
2.14 Phosphorylation of HPV-16 E7 followed by time-resolved experiments 55
2.15 Sample preparation and their analysis/assays 56

Chapter 3. Results 58

3.1 Structural and dynamic characterization of the molecular hub E1A from human adenovirus 62
3.2 Toward the real-time monitoring of HPV-16 E7 phosphorylation events 77
3.3 The structure and dynamics of HPV-16 E7 described by NMR-guided metadynamics 97

Chapter 4. Conclusions 121

Report of activities 125

Appendices 127

References 133

List of Abbreviations

CBP	CREB Binding Protein
CDK	Cyclin-Dependent Kinase
CKII	Casein Kinase II
CR	Conserved Region
DTT	1, 4-Dithiothreitol
HAdV	Human Adenovirus
HPV	Human Papillomavirus
IB	Inclusion bodies
IDP	Intrinsically Disordered Protein
IDR	Intrinsically Disordered Region
IPTG	β-D-1- thiogalactopyranoside
LB	Luria-Bertani medium
MORFs	Molecular Recognition Features
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effect
OD	Optical Density
pRb	Retinoblastoma Tumor Suppressor Protein
PTM	Post-translational Modification
SLiMs	Short Linear Motifs
SSP	Secondary Structure Propensity
TBP	TATA-box Binding Protein

Chapter 1

Introduction

Viruses are small obligate intracellular parasites involved in a huge variety of diseases going from simple infections to severe complications that may cause ultimately death. Due to the limitations imposed by their genomes, which are not large enough to encode all biomolecules required to reproduce themselves, viruses are absolutely dependent on the host cell functions and complex networks of interactions [1].

Viruses and cancers have shown to be closely related. It has been shown that approximately 12% of human cancers worldwide are strictly connected to viral infection, with more than 80% of cases occurring in the developing world [2,3]. Cancer incidence is much lower than virus prevalence in human population. It shows that oncoviruses may be necessary but they are not a determinant factor for cancer development [4]. The infection process as well the oncogenic virus potential is directly correlated to the interactions of the virus and host cell proteins. Hence, the knowledge about viral proteome and about protein interactions is crucial for understanding viral infection mechanisms and to identify target proteins involved in host cell regulation.

Eukaryotic and non-eukaryotic cells frequently make use of alternative splicing to express multiple proteins from one gene, increasing the diversity of proteins to be expressed. With the same purpose, viruses often use the host machinery to obtain alternative splicing of their mRNA in order to compensate their limited genome architecture. Another way found by organisms to increase diversity is the construction of multi-target (hub) proteins [5].

Hubs are, by definition, proteins able to bind to a large number of other proteins, creating highly connected nodes of links inside complexes protein-protein interaction networks [6]. The ability to constitute a high level of intermolecular

interactions, as observed in hub proteins, is less compatible with globular proteins or globular domains features, which, in many cases, bind with high-affinity to a specific target.

The first evidences that non-structured segments of proteins could play important roles in protein functions appeared in the literature more than 20 years ago [7,8]. These evidences require an extension of the structure function paradigm, originated by the "lock and key" hypothesis formulated by Emil Fischer in 1894 which affirms that a three-dimensional structure is a prerequisite for a functional protein [9]. Nowadays it is clear that hub proteins tend to be more disordered than other proteins [10]. In this context, intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) occupy a very important place. In fact, IDPs/IDRs frequently interact with or act as hubs [11,12].

IDPs/IDRs are characterized by the lack of a three-dimensional stable fold and by a high flexibility, existing as dynamic conformational ensembles. All the structural properties of IDPs/IDRs are the result of their amino acid sequence which has a low content of hydrophobic residues, respect to folded proteins [13–15] while it is in general rich in aspartic acid, glycine, lysine, arginine, serine, glutamine, proline and glutamic acid. Most of these residues are polar, often charged, and commonly found on the surfaces of ordered proteins [16,17]. This explains why IDPs are unable to fold spontaneously into a stable well-defined three-dimensional structure [16–19]. Although the term IDPs and IDRs is now used to refer to highly flexible and disordered proteins, not all of them are in fact entirely disordered. Several proteins that contain disordered sequences in combination with more structured globular domains, resulting in highly heterogeneous systems are still called IDPs [20].

IDPs/IDRs play important roles in regulatory functions and signalling pathways being crucial in so many cellular processes including regulation of transcription, translation and cell cycle control [21–24]. Thanks to their structural plasticity and flexibility, IDPs/IDRs can associate with several unrelated partners with a high specificity and a moderate affinity, acting as sensitive "on-off" switches [25]. Moreover, IDPs present accessible sites for post-translational modification being easily modulated by interaction with partners and more generally by environmental

conditions [19,20] . In some cases protein-protein interactions can be mediated by short segments of linear sequences named molecular recognition features (MORFs) or short linear motifs (SLiMs), two additional ways to maximize their interaction capabilities [11]. These features are extensively used by viruses that need to hijack the host cell, using the host machinery to work in their own favour. In fact, many viral proteins are found to be intrinsically disordered. The percentage of distribution of ID in organisms estimates that eukaryotes exhibit 33% of their proteins with such long predicted regions of disorder. Bacteria and *archaea* are expected to have 2-4% of IDPs/IDRs in their proteomes[25]. On the other hand viruses have the widest spread of the proteome disorder content, ranging from 7.3% to 77.3% depending of the virus type [26,27].

Among the several viruses that make use of intrinsically disordered proteins are human papilloma virus (HPV) and human adenovirus (HAdV) [28]. Several proteins encoded by these viruses play similar roles in several human pathways. Good examples are E7 from HPV and E1A from HAdV. Both proteins are predicted to have heterogeneous structural and dynamic properties and despite they are related to different diseases, they share several similarities on the primary sequence and its organization. They also contain two CXXC motifs which are expected to bind zinc and form a more structured, although minimal, protein domain. Moreover, E7 and E1A share common interaction partners and some of these interactions are modulated by post-translational modifications. These and various other aspects, addressed in Sections 1.1 and 1.2, encouraged us to focus on the structural and dynamic characterization of both proteins.

Considering the high level of distribution of IDPs/IDRs in organisms, in particular for viruses and their high biological importance, it is necessary to contribute to the structural characterization of these important classes of proteins.

X-ray crystallography is one of the major techniques to access high-resolution information on large biomolecules. A great number of protein structures have been elucidated through X-ray crystallography. However, IDPs and IDRs usually fail to give crystals and/or to coherently scatter X-rays due to their highly flexible properties [29]. In some cases the NMR structure in solution reveals also dynamic features that are more difficult to observe by crystallography [30,31]. Indeed, results obtained by X-ray

and NMR analyses of the same protein revealed that the structures might be quite different. For X-ray crystallography analysis the protein needs to be in the crystalline state, which consequently reduces its natural flexibility. In fact, the crystal lattice favours only one of the several conformations that a protein could adopt in solution [32].

NMR spectroscopy is the unique technique to access atomic resolution information on structural and dynamic properties of IDPs/IDRs in solution. Indeed it is a powerful tool to study interactions and naturally occurring reactions such as post-translational modifications, at high resolution.

1.1 E7 protein from human papillomavirus

The human papillomavirus (HPV) family comprises more than 100 different viruses that infect the basal layers of mucosa and the stratified epithelium. HPV has been found in over 90% of cervical cancers, as well as in other carcinomas [33] .

Based on their pathogenicity, HPVs are classified in two groups: The "Low-risk" (LR), which may cause benign warts and the "High-risk" (HR), related to cancer development [34,35]. Within the HR group, HPV-16 and HPV-18 are responsible for almost 70% of all cervical cancers, the third most common cancer in women worldwide [36,37]. In particular, HPV-16 is considered the most carcinogenic type and it is more often found in cervical cancer than HPV-18. Furthermore, HPV-16 is uniquely associated to oropharyngeal tumours [34,35,38].

The infection by HPVs requires the access by the virus to the epithelia basal layers, usually allowed by the presence of small wounds in the epithelium. Once there, the virus particles use the host cell machinery to amplify their genome. After a while, the replication takes place in the differentiating keratinocytes, leading to their abnormal cell proliferation [39]. During this infection process, the HPV's genome coordinates the expression of eight viral proteins: L1 and L2, structural proteins, and the non-structural proteins E1, E2, E4, E5, E6 and E7 (Figure 1.1).

L1 and L2 are the major and minor capsid protein, respectively [40]. E1 and E2 are directly involved in the viral DNA replication, regulating the early transcription [41,42]. E4 accumulates in the cell during the viral genome amplification, promoting the differentiation-dependent productive phase of the viral life cycle [43–45]. E5 supports late functions and modulates cell signalling, allowing cell cycle progression; E5 is also important to disrupt the MHC class II maturation [46,47].

E6 and E7 work as oncoproteins in the HR HPVs, interacting with the cell cycle regulators p53 and retinoblastoma tumour suppressor (pRb), respectively [48,49]. Despite E7 and E6 work in cooperation to haywire the cell cycle, HPV-16 E7 alone is sufficient to immortalize cells, showing its high importance in the tumorigenic process [50].

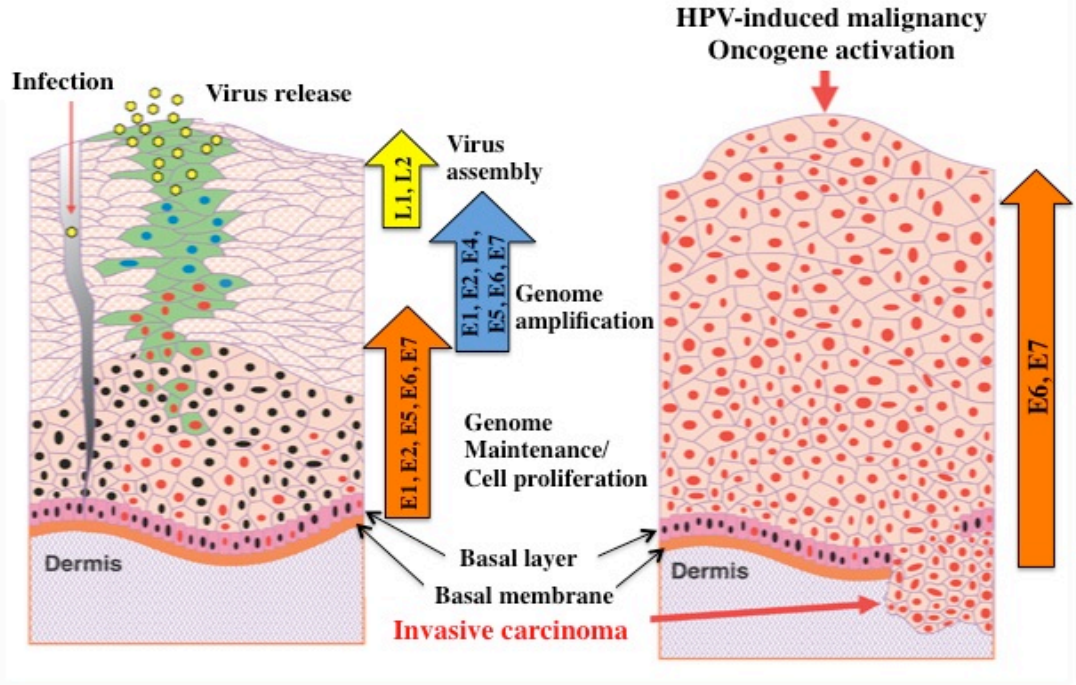


Figure 1.1. Schematic representation of the HPV infection. On the left the different layers of the epithelium are shown. The virus particles infect the basal layer of the keratinocytes establishing the viral genomes and expressing the early viral proteins E1, E2, E5, E6 and E7 (orange arrow). During this phase the viral genome is replicated in synchrony with the host DNA replication. After cell division, keratinocytes undergo differentiation. In infected cells, E6 and E7 deregulate the cell cycle control, pushing differentiating cells into S phase (blue arrow). In the late-phase, the viral capsids are assembled through the expression of L1 and L2 proteins, allowing the release of new virus copies. Adapted from Thomas *et. al*, 2008 [39,51].

E7 is a small protein with almost 100 amino acids. Its length and amino acid composition may be different among the different types of HPV, influencing its affinity with partners and, consequently, its oncogenic potential. Within the HPV family, E7 shows the presence of three conserved regions (CR), named CR1, CR2 and CR3 (Figure 1.2) [52,53].

As a result of bioinformatics predictions, CR1 and CR2, located in the N-terminal part, are expected to be intrinsically disordered, while CR3 is expected to be more structured [54–56]. The first two domains (CR1 and CR2) have been described to interact with many target proteins. The CR2 presents a specific LXCXE motif which targets the main partner of E7, pRb [49,57]. E7 proteins from HPV strains of the LR group have reduced or no affinity for pRb suggesting that the binding to pRb is essential for the ability of E7 to transform cells [58]. Once bound to pRb, the

transcription factor E2F is displaced, resulting in abnormal DNA transcription and forcing the cells to go in S phase [54,59].

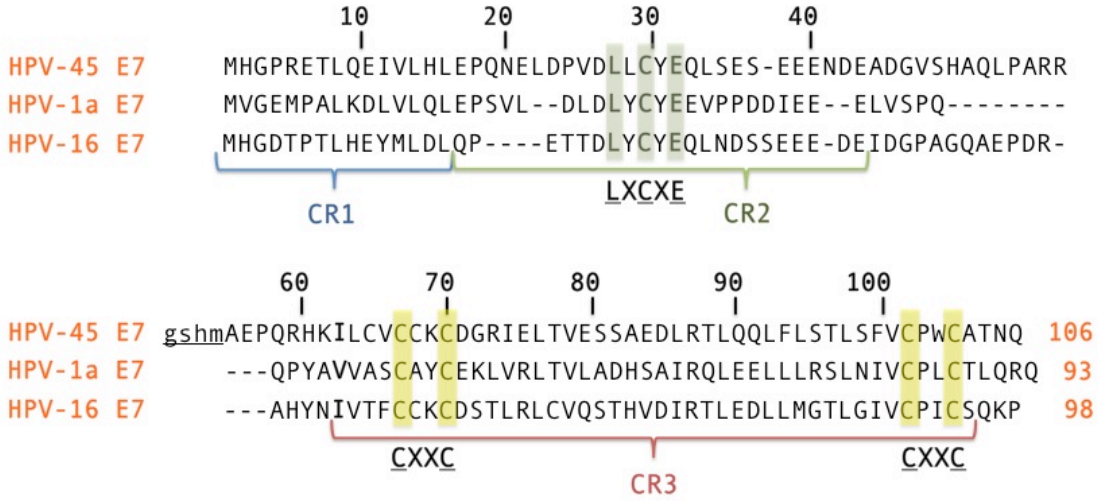


Figure 1.2. The amino acid sequence of E7 from the LR HPV-1a and the HR HPV-45 and HPV-16 and its conserved regions CR1, CR2 and CR3. The length of the protein for each HPV type is shown on the right. The pRb binding motif LXCXE is highlighted in green. The zinc-binding site is highlighted in yellow. Adapted from Roman and Munger, 2013 [53].

In fact HPV-16 E7 acts as a hub-protein, interacting with several partners involved in the cell cycle control and apoptosis (Figure 1.3). Besides the interaction with pRb, E7 interacts also with its family member p107 and p130, important tumour suppressors responsible for the G_1/S check point of the cell cycle [38,60,61]. E7 also associates with other proteins involved in cell proliferation like cyclin A- and cyclin E/cyclin-dependent kinase-2 (A/CDK-2 and E/CDK-2, respectively) and p21 and p27 protein [62–64]. E7 binds also to the transcription factors TATA box-binding protein (TBP), p300/CBP and E2F, completely deregulating the normal cell cycle [65–67]. Moreover, E7 presents post-translational modifications (PTMs) sites, in particular for phosphorylation, that allows the modulation of some of the most important interactions of E7 by increasing its affinity for specific partners [68–71].

CR3 is also a very important domain for the biological activity of E7, including the interaction with pRb (Figure 1.3)[72]. It presents two CXXC zinc-binding motifs separated by 29 amino acids. The integrity of the zinc-binding domain is very important for the intracellular stability and for the aggregation state of E7 [73–75]. In solution, E7 can be found as monomer in equilibrium with dimeric, tetrameric and

oligomeric forms but under physiological conditions the majority is found as a dimer[75–78].

The structural and dynamic characterization of E7 has been achieved only through short constructs. Till now, the whole protein had not been characterized. It fails to produce crystals and the complete characterization through NMR could not be achieved, despite documented attempts to obtain it [77]. The NMR characterization is available for CR1 and CR2 of HPV-16 E7, revealing the intrinsic disorder and the high flexibility of these domains [56]. CR3 from LR HPV-1a E7 was characterized by crystallography and the NMR solution structure is available for HR HPV-45 [54,77]. For both structures, CR3 consists of two β-strands and two α-helices mediating the formation of a homodimer. The E7's topology is unique, not present in any other known zinc-binding protein. Since HPV-45 E7CR3 and HPV-16 E7CR3 are closely related and their sequences share almost 43% identity and 73% of similarity [78], HPV-45 E7 structure could be used as a template for constructions of a HPV-16 E7 model through homology methods (Figure 1.3).

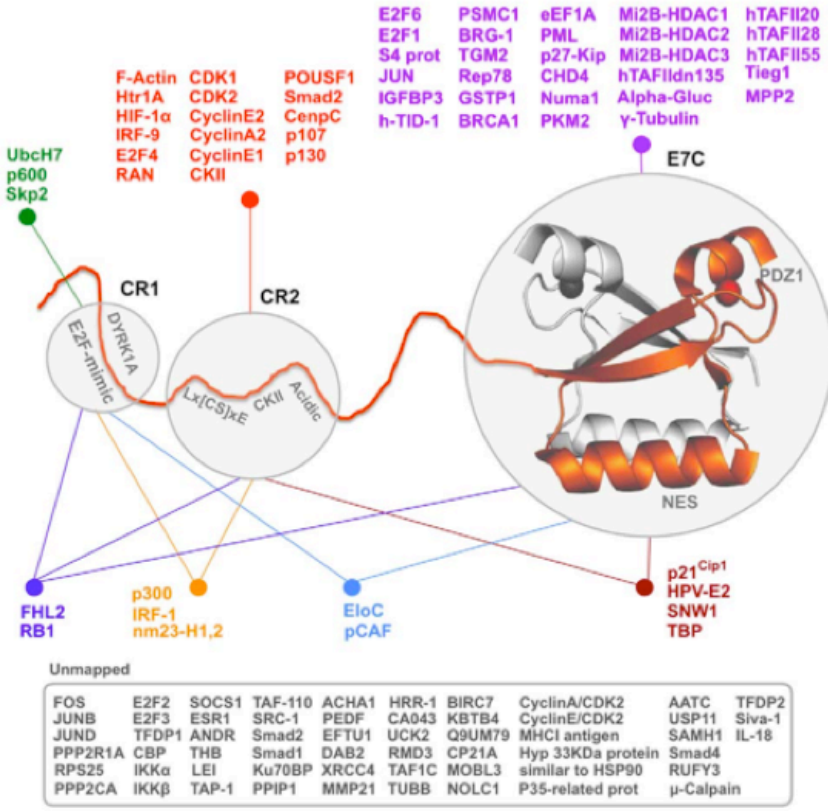


Figure 1.3. Schematic representation of HPV-16 E7 and its partners, including the structure location of the conserved regions CR1, CR2 and CR3. Adapted from Noval *et al.*, 2013 [79].

Despite the importance of HPV-16 E7, its CR3 domain was not yet characterized. The atomic resolution structural and dynamic information of this domain is extremely important to better understand the interactions in which E7 is involved as well to figure out the basic differences between the different E7 variants in the HPV family.

1.2 Human adenovirus E1A proteins

Adenoviruses (AdVs) are members of the *Adenoviridae* family. These are non-enveloped medium–size viruses (about 90 nm), with an icosahedral capsid packing a linear double-stranded DNA of 34 - 48 Kb (Figure 1.4) [80–82].

Since their isolation from human adenoids, AdVs have been extensively characterized and classified according to their DNA sequences and to their agglutination properties [83]. Nowadays over 100 distinct serotypes have been isolated from a broad range of hosts including birds, reptiles and amphibians [81]. More than 50 serotypes were isolated from humans and are related to an array of clinical diseases like respiratory infections, conjunctivitis and gastroenteritis [84–87].

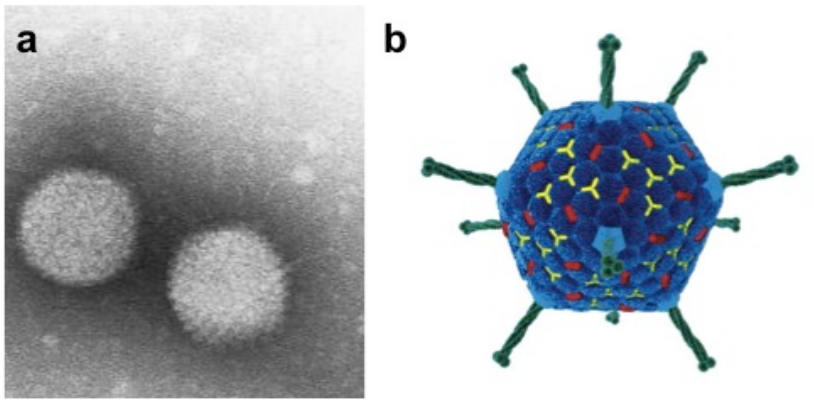


Figure 1.4. (a) Electron microscopy of a Human adenovirus particles and (b) a three-dimensional representation of the icosahedral virion [80,82] .

The discovery that some human AdVs (HAdVs) induce undifferentiated sarcomas in new-born hamsters and rodent cells stimulated the characterization of the many aspects of its lifecycle. These findings established HAdV as a model system to study the molecular mechanisms controlling cell cycle progression and apoptosis [88–90].

Among the HAdVs, the closely related serotypes HAdV-2 and HAdV-5 are the most extensively studied in particular for vaccine development and clinical trials. Their genome is approximately 36 kbp encoding six early transcription units (E1A, E1B, E2A, E2B, E3 and E4) and one predominant late promoter that generate five late

transcripts (L1-L5) [91,92]. Each transcript can raise multiple mRNAs by alternative splicing.

E1A is the first gene to be expressed after HAdV infection. Initial studies demonstrated that the expression of E1A from HAdV-5 is sufficient to transform primary rodent cells [93] and that E1A is the main factor that determines whether the transformed cells are tumorigenic in immunocompetent animals [94]. However, controversially, it was later demonstrated that E1A products act as a tumour suppressor in specific human cells [95]. These findings pose fundamental questions on the molecular origins of this apparently controversial behaviour and demand further investigation on the molecular determinants of E1A function [96].

Through alternative splicing HAdV E1A gene gives rise to five different mRNAs named 9S-13S, according to their sedimentation coefficients. The products of 9S, 10S and 11S mRNAs are expressed later and their functions are not well understood yet. Instead, the 12S and 13S mRNAs code for two main proteins expressed right after the HAdV infection: a 243 and a 289 amino acid protein (E1A243 and E1A289, respectively).

The sequence analysis of E1A243 and E1A289 from different serotypes of HAdV reveals the presence of four conserved regions (CRs) called CR1-CR4, that are connected by less conserved segments. The major difference between E1A243 and E1A289 proteins consists in the presence of the CR3 domain (46 residues) in E1A289 but not in E1A243 (Figure 1.5) [97,98].

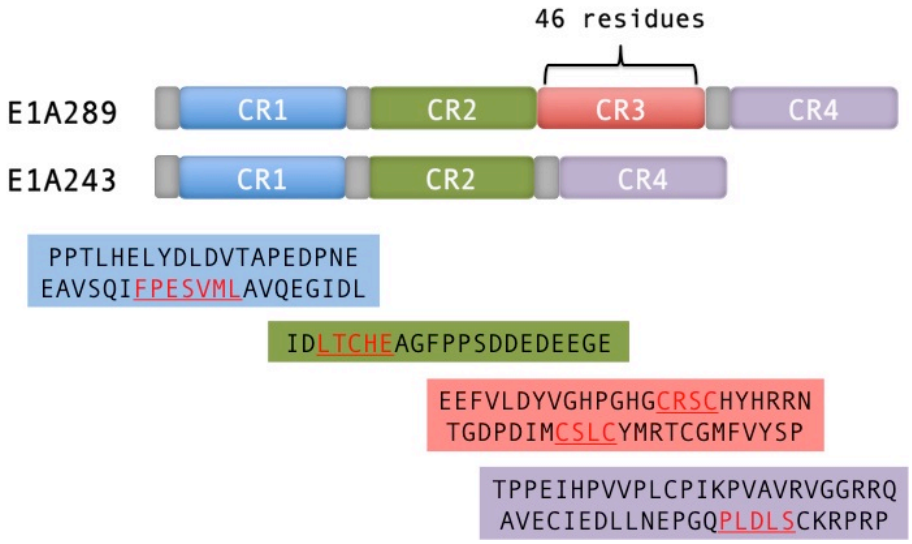


Figure 1.5. Schematic representation of the alignment of E1A289 and E1A243 from HAdV-2/5. It shows the disposition of the conserved regions (CRs) 1-4 for both proteins. The most important motifs for the main interactions and structure stability of E1A are highlighted. Adapted from Chinnadurai, 2011 [99].

Bioinformatic analysis using disorder predictors suggests that a large portion of the protein has a high propensity to be disordered, although several regions showing tendency to be more ordered can be identified (Figure 1.6) [100].

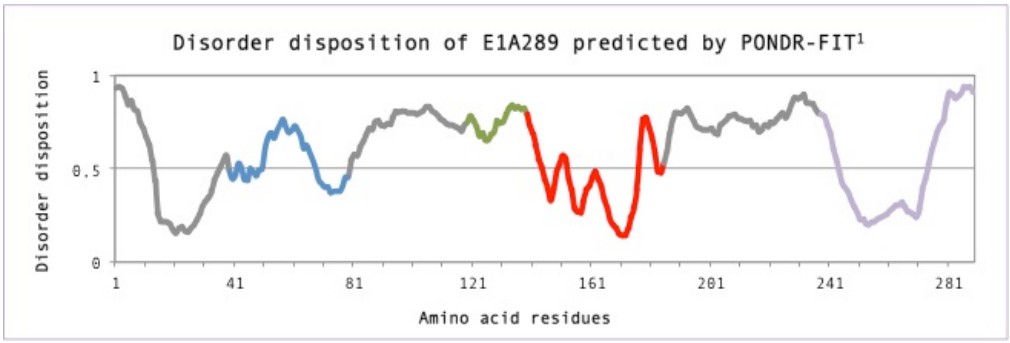


Figure 1.6. Disorder disposition of E1A289 predicted by PONDR-FIT. The colors represent the conserved regions present in the protein: blue for CR1, green for CR2, red for CR3 and light violet for CR4. The analysis shows that E1A289 is expected to be a very heterogeneous structure where the initial N-terminus (residues 1-40), the CR3 and CR4 would be the well-structured parts of the protein [101].

Both E1A243 and E1A289 (E1A) have been reported to alter or inhibit the function of several cellular proteins. The conserved regions are responsible for important interactions with host proteins, often supported by other parts of the protein not necessarily well conserved [28,102].

The CR1 domain is reported to cooperate with the N-terminal part (first 41 residues of E1A) during the interaction with transcriptional co-activators such as p300/CBP through the FPESVML sequence (residues 66 - 72)[103] and with the member of the chromatin remodelling proteins, p400/TRRAP (the transactivation/transformation-domain-associated protein) via the ELYD sequence (residues 45 - 48) [104]. E1A might sequester TRRAP, and lead to a loss of transcriptional activation by factors that recruit these complexes. However, the molecular mechanisms and the consequences of this interaction are still not clear [105,106].

The most important and best-characterized interactions of E1A are related to cell cycle control partners and transcription factors. Both E1A243 and E1A289 bind pRb through an LXCXE motif present in the CR2 domain (residues 122 -126) in a similar way as E7 from HPV [107,108,60]. However, in case of E1A, this interaction is anchored by the CR1 domain, making two distinct points of contacts [109].

The CR2 may interfere with crucial pathways in the host cell. It mediates the interaction with the potent transcriptional repressor BS69, through the PXLXP motif (residues 113 - 117). This PXLXP sequence is not highly conserved, being expressed in some E1A from few serotypes of HAdVs [98]. The interaction requires also the CR3 domain. As consequence, E1A243, which misses CR3, has much lower affinity for BS69 [110–112]. It interacts with proteins from the small- ubiquitin-related modifier (SUMO) family, which regulate several cellular processes such as nuclear transport, transcription, chromosome segregation and DNA repair [113,114]. Moreover, CR2 binds to the S2 component of the 19S regulatory complex of the 26S proteasome interfering with its activity [115]. The 26s proteasome is responsible for balancing the amount of proteins inside the cell by degrading unneeded or damaged proteins. One of these proteins is the tumor suppressor p53. The interaction of E1A with S2 contributes to the stabilization of p53 and consequently induces apoptosis processes [116].

The CR3, present only in E1A289, contains two CXXC motifs (residues 154 – 157 and 171 – 174) expected to form a zinc-binding domain [117]. It works as a potent transcriptional activator domain, being thus essential for activating viral early gene expression [118–120]. E1A289 interacts, with the transcription factors like TATA box-binding protein (TBP) and with a component of the mediator adaptor complex MED23, through specific residues on the CR3 [121,122]. The interaction is important to nucleate the transcriptional pre-initiation complex [122]. Nowadays, all deletions or mutations within CR3 resulted in decrease or in the complete loss of transcriptional activation function [123–126].

The 26S proteasome was shown to interact with the CR3. 26S is the major non-lysosomal proteolytic machinery in eukaryotes. It is a complex composed by two large macromolecular complexes: the proteolytic particle, 20S; and the regulatory complex 19S [127]. CR3 was shown to be able to interact with 19S independently of 20S complex via the residues 169 -188. The 20S proteasome subunit has also been shown to interact with CR3 independently. Chemical inhibition of the 26S proteasome also represses CR3-dependent activation of transcription, suggesting that the proteasome directly controls E1A-dependent transcriptional activation.

CR4 domain of E1A is 48 amino acid residues predicted to be structured, as noted previously in the Figure 1.6. This C-terminus region is required for oncogenic transformation of rodent cells, cooperating with E1B [128]. Indeed, it seems to be required for the activation of transcription of specific genes in the epithelium, inducing the mesenchymal to epithelial transition [99]. However, paradoxically, it seems to suppress cell transformation by interaction with the activated Ras, leading ultimately to a negative modulation of tumorigenesis and metastasis [129,130].

Despite decades of extensive studies, the function of CR4 as well its binding partners are not well characterized. Among the few known proteins that interact with this domain, the binding to the E1A C-terminal binding protein (CtBP) family via the PXDLS motif (residues 279 to 283), is the best characterized [130,131] . CtBP is a transcriptional corepressor that dimerizes after NADH binding and, ultimately, drives the formation of a silencing complex [132,133]. E1A competes with the cellular targets of CtBP, inhibiting its activities and consequently leading to a derepression of

transcription of specific repressed promoters. Hypothetically, E1A can use the CtBP to access those repressed promoters, then recruit coactivators to turn on the gene transcription [134].

Despite the biological importance of E1A243, E1A289 and its CR3 domain their structure and dynamic properties remain still undetermined.

1.3 Post-translational modifications

While the human genome is estimated to comprise about twenty-five thousand genes, the total number of proteins is estimated to be over one million [135–137]. This feature is possible thanks to the genomic recombination, transcription initiation at alternative promoters, differential transcription termination, and alternative splicing. In this way, single genes can generate different mRNAs and thus encode multiple proteins [138,139]. Moreover, proteomic diversity is vastly increased by post-translational modifications (PTMs).

PTMs are chemical modifications of specific protein sites that play a key role in functional proteomics by influencing numerous properties of proteins including enzymatic activity, protein interactions with other biomolecules as well their subcellular location [140]. These chemical modifications occur through different mechanisms in response to different conditions in a time- and signal-dependent way, allowing the cell to regulate a range of functional/structural responses that are crucial for several cellular processes.

Fast cellular signalling responses usually require the quick access by modifying enzymes to specific recognition sites (PTM sites), encoded into the primary sequences of proteins [141]. In this context, the naturally solvent exposed amino acids of IDPs/IDRs offer advantages for the easier occurrence of PTMs [142–145].

More than 200 different types of PTMs have been described in the literature, ranging from small chemical modifications such as phosphorylation and acetylation to the addition of proteins or sugars by ubiquitination and glycosylation, respectively [146]. The combinatorial use of these different types of PTMs in a transient and reversible way allows the creation of a complex signalling network. The Table 1.1 shows the most common PTMs in eukaryotic cells.

Some reactions also provide different forms, for example lysines may be mono-, di- or tri-methylated. This variety of different states for a single amino acid further increases the diversity and plasticity of protein functions.

Table 1.1. Examples of the most common post-translational modifications in eukaryotic cells.

Modification	Added group	Modified residues	Example of Enzymes
Phosphorylation	Phosphoryl	Ser, Tyr, Thr, His	CKII, PKA, TK, CAMK
Acetylation	Acetyl	Lys	HAT, Nat family
Methylation	Methyl	Lys, Arg	Histone methyltransferases
Nitrosylation	Nitric oxide	Cys, Met	SOD, NOS
Glycosylation	Glycosyl	*N*-Glyscosylation: Asn, Arg, N-terminus	GlcNAc-T1 and Mann II
		O-Glycosylation: Ser, Thr and amino acids close to Tyr phosphorylation sites	
Ubiquitination	Ubiquitin	Lys, N-terminus, Cys, Thr, Ser.	E2, E3

1.3.1 PTMs of E1A proteins

E1A presents several possible sites for phosphorylation in particular within CR1 and CR2. The major sites were mapped to Ser89 and Ser219 that are substrates of Cdk-like protein kinases [147,148]. The phosphorylation of Ser89 of both E1A proteins improves the efficiency of interaction with pRB in disrupting the complex E2F-pRB [148] and their ability to mediate cell transformation [149–151]. On the other hand, mutations on Ser129 do not have pronounced biological effects [152]. Another residue phosphorylated near the N-terminal part is Ser96 by a yet unidentified enzyme. The phosphorylation of this serine seems to be regulated by the phosphorylation of Ser89 [149–151].

An additional and very important phosphorylation site is Ser132, which is phosphorylated by CKII. Ser132 is located in the CR2 domain, near the pRB binding site found also in HPV-16 E7. In E1A289, this serine residue is just upstream of CR3, the important domain involved in the regulation of early viral gene expression [153].

The phosphorylation of Ser132 residue in E1A reduces its apoptosis-inducing ability, thus increasing its transforming potential. This phosphorylation may enhance the binding of pRB and related proteins. In the case of E1A289, it also reduces cytotoxicity, thus increasing its transforming activity [153] .The replacement of Ser132 by alanine or glycine reduces the efficiency of these interactions. However, these mutations do not kill the transformation ability of E1A proteins in primary rat cells when E1A is in cooperation with E1B protein [153].

Within the C-terminal part of E1A289 two that can be phosphorylated include Ser185 and Ser188. Studies demonstrated that at least one of these two sites, is phosphorylated in vitro by purified mitogen-activated protein kinase (MAPK), and both are hyperphosphorylated in cells which express a constitutively active form of MAPK kinase [154]. Another PTM within the C-terminal part is the acetylation of Lys285 which disrupts the interaction with the transcriptional co-repressor CtBP [155].

Another important but not well-characterized PTM of E1A is the phosphorylation of the CR4 domain by the dual-specificity tyrosine phosphorylation-regulated kinase 1 and 2 (DYRK1A and DYRK1B, respectively). DYRK1A/1B are kinases involved in regulating cell proliferation and differentiation survival. These interactions of E1A with DYRK1A/1B seem to overlap with the CtBP binding site [156].

1.3.2 PTMs of E7

The sequence analysis of E7 predicts PTMs sites, in particular phosphorylation sites spread throughout the whole protein. E7 contains consensus sequences for glycosylation at Asn29 and two sulfation sites at Tyr23 and Tyr26. However, till today, only phosphorylation reactions have been reported. HPV-16 E7 is phosphorylated at Thr5 and Thr7 by DYRK1A. This phosphorylation interferes with the degradation of E7 by 26S proteasome and thus increases its transforming potential by extending the half-life of E7 within the cells [157]. The phosphorylation of Ser71 is

also reported to occur *in vivo* by an unidentified kinase and the biological effects of this modification remain still unknown [69].

The most important phosphorylation of E7 is performed by CKII at Ser31 and Ser32 within a conserved motif present also in E1A proteins, as mentioned before [69]. In this way, CKII represents a very important modifying enzyme present in human cells required to enhance the oncogenic potential of E7 and E1A proteins.

As mentioned before, the CR2 domain of both HPV-16 E7 and E1A proteins plays a critical role in its oncogenic function by interaction with pRb via its LXCXE motif. On E7 this highly flexible domain contains two phosphorylation sites (Ser31 and Ser32) in a recognition site for CKII (-N-D-**S-S-**E-E-E-D-E-) very close to pRb binding site [158,159].

It has been shown that E7 is differentially phosphorylated during the cell cycle at G1 phase by CKII and at S phase by enzymes not yet identified [69]. The phosphorylation of both Ser31 and Ser32 by CKII is shown to enhance not only the interaction with pRb but also with CBP [70] and TBP [65,69], contributing to E7's transforming activity. Indeed, the phosphorylation of HPV-18 E7 on Ser32 and Ser43 has been shown to be critical to promote S-phase entry [71]. In a similar manner, mutations of both serines in the consensus sequence for HPV-16 E7 (Ser31 and Ser32) affect its transforming ability [159,160]. Interestingly, the replacement of Ser31 and Ser32 by aspartic acid residues, creating a negative charge mimicking the phosphorylated residue is able to restore the transforming activity of E7 [159,160]. In addition, the level of E7 phosphorylation and transformation activity is lower in low-risk HPV types than in high-risk HPV types [159]. All these results show that the phosphorylation of E7 by CKII is critical for the enhancement of its oncogenic ability and for the development of HPV-associated neoplasia.

1.4 Casein Kinase type II and its requirements to perform phosphorylation

While some protein kinases like the extracellular-signal-regulated kinase (ERK) and mitogen-activated protein (MAP) kinase are more specific, phosphorylating only one or two distinct protein targets, other kinases show to be much less selective, phosphorylating hundreds of distinct proteins. This is the case of the protein casein kinase II (CKII).

CKII is a small family of closely related serine/threonine kinases involved in multiple functions including signal transduction, cell division and proliferation. It is overexpressed in malignant tumours representing thus a molecular target for drug design [161]. Studies have demonstrated the ability of CKII to phosphorylate also tyrosine under special circumstances, showing the plasticity of the recognition site of this kinase [162]. CKII is present in the cytoplasmic and nuclear compartments, which is consistent with its wide spectrum of physiological targets, and its multiple functional activities within the cell [163–165]. It is independent of second messengers or phosphorylation for activation and is able to use GTP besides ATP as a phosphate donor.

In terms of structure, CKII is a tetrameric enzyme consisting of two catalytic (α) and two regulatory (β) subunits [166,167]. While only one form of regulatory subunit is known, the catalytic subunit may exist as different isoforms in many organisms [168–170]. In humans three catalytic isoforms have been found (α, α' and α'') [171,172] . All these three catalytic isoforms are very similar with the exception of their unrelated C-terminal domains. The catalytic isoforms exhibit a great functional overlap. Evidence suggests a functional specialization of the different CKII isoforms in mammals, mice and yeasts [173]. However, due to their high similarities it is not possible to determine which isoform contributes more to a specific reaction.

CKII is an acidic-directed kinase i.e. it requires acidic residues like Asp or Glu close to the modified Ser/Thr (serine being preferred over threonine) [162]. Acidic amino acids are more strongly selected at the C-terminal position of Ser/Thr than at their N-terminal position. Although the presence of acidic residues on the N-terminal side of

the serine/threonine is not required for recognition, it can improve the kinetic constants of phosphorylation by CKII [174].

Some positions of the acidic residues are preferred by CKII for substrate recognition and for kinetic improvements. The presence of an acidic residue at *n+3* position (where *n* is the phosphorylatable amino acid, serine or threonine) is crucial for substrate recognition by CKII. The second most important position is *n+1* [175–179]. The presence of an acidic residue at this position in fact is not required for substrate recognition but it can considerably improve the Michaelis-Menten constant (Km). On the other hand, peptides presenting the position *n+2* occupied by an acidic residue or by an arginine are equally phosphorylated, suggesting that an acidic residue at this position is not important for substrate recognition or kinetic improvements. The minimal structural requirements for substrate recognition by CKII is represented on Figure 1.7.

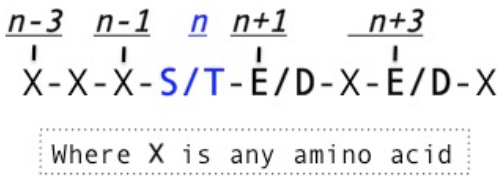


Figure 1.7. Minimum substrate requirements for CKII phosphorylation. CKII recognizes and phosphorylates serines (S) and threonines (T) in acidic regions. The presence of acidic residues at *n+3* and *n+1* position, respect to the S or T is required.

HPV-16 E7 and E1A fulfil these minimum requirements for CKII phosphorylation (Figure 1.8). As it is possible to see, Ser132 of E1A and Ser32 of HPV-16 E7 fulfil both *n+3* and *n+1* requirements. Instead, Ser31 of E7 fulfils only the *n+3* requirement. For this residue, by the sequence analysis, one could expect a lower phosphorylation rate respect to its neighbour, Ser32.

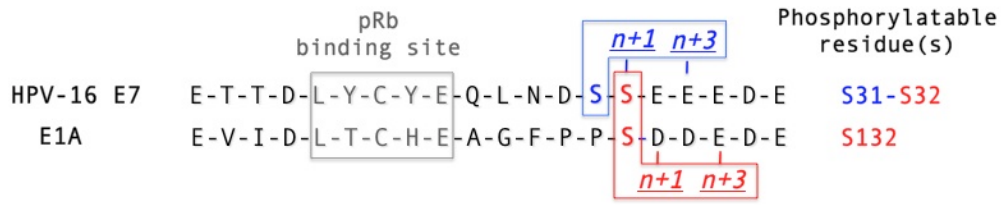


Figure 1.8. The specific recognition site for CKII found in HPV-16 E7 and in E1A. In Both proteins the minimum sequence requirement for CKII phosphorylation is present. In HPV-16 E7 Ser31 fulfils only the *n+3* requirement, while Ser32 in HPV-16 E7 and Ser132 in E1A fulfil the *n+1* and *n+3* requirements

1.5 Studying PTMs by NMR techniques

As mentioned in Section 1.3, most part of PTMs occurs by the covalent addition, to the side-chains of proteins, of small chemical groups, such as phosphoryl groups, acyl and alkyl chains and sugars. These modifications induce specific changes in the chemical environments of the modified residues. More specifically, phosphorylation results in a large downfield shift for protons of the amide resonances ($\Delta\delta \approx 0.5/1.5$ ppm) and thus they can be readily detected at atomic resolution through NMR spectroscopy. Indeed, NMR is the only tool that allows studying PTMs at atomic resolution on a mechanistic and structural perspective, by following these modifications in a time-resolved manner either *in vitro* [180] *or in vivo* [181].

Another advantage of studying PTMs by NMR is the possibility to follow reactions at multiple sites. Distinct modifications that may occur in parallel (i.e. phosphorylation and acetylation) can be easily identified by NMR; this would be challenging for most analytical methods, in particular for mass spectrometry (MS), which is based on proteolytic processing routines and peptide fragment-based detection. As illustrated in Figure 1.9, MS could generate identical fragments from two different PTMs that cannot be distinguished by MS without elaborated identification processes of MS/MS fragmented peptides. In contrast, since the presence of one PTM influences the chemical shifts of the modified amino acids and of the adjacent ones, the corresponding NMR peak patterns can be unambiguously identified whether both PTMs are present on the same, or on different molecules.

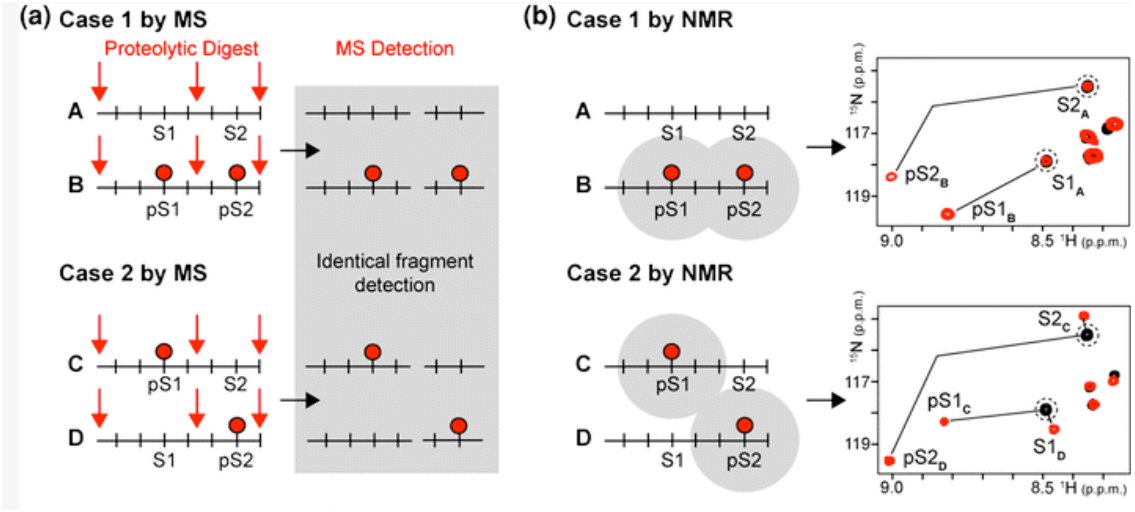


Figure 1.9. a) Schematic illustration of the information provided by MS and NMR on PTM sites. By MS analysis (a), identical pairs of PTM/peptide fragments are generated from two different PTM distributions (case 1 and case 2). On the other hand, by NMR (b), these PTMs could be individually identified, since each modification influences the chemical environment, and hence resonance frequency, of the respective other site. Adapted from Theillet, F *et al.*, 2012 [182].

The quantitative nature of NMR spectroscopy is another feature that makes it particularly interesting for PTM studies. The reaction occurring can be followed by time-resolved experiments (acquisition of a series of uninterrupted bidimensional spectra). The changes in NMR signal intensities of modified and unmodified substrates residues are related to changes in their concentration and can be measured through simple NMR signal integration. With this, kinetic reaction parameters can directly be extracted [183]. These measurements are particularly useful in providing additional mechanistic insights into stepwise PTM reactions, in particular for multistep reactions where the modifications happen in a defined order [181].

Conformational changes reported as a direct consequence of individual PTMs are immediately assessed by NMR, by observing additional resonance peaks displacements that may occur on other regions of the protein (distant from the specifically modified residues). The resulting chemical shift changes often increase spectra complexity, as they no longer involve only the modified residues alone. Protein phosphorylation in particular has been reported to promote conformational changes including modulations in α-helix stability or C-terminal destabilization [184–188].

Phosphorylation reactions are the most abundant PTM in eukaryotes and typically involve different kinases. In fact, phosphorylation and dephosphorylation act as molecular switches regulating many cellular processes [189,190]. The human genome encodes several hundred distinct kinases and one third of all cellular proteins appear to be phosphorylated [191,192]. In many cases different reactions are performed in the same protein/domain, resulting in several combinations of possible protein states that play a key role in signalling processes [193,194].

Serine residues are the most common amino acid phosphorylated in eukaryotes, whereas threonine and tyrosine phosphorylation are less common [195,196]. These amino acids present a nucleophilic hydroxyl group (–OH) that attacks the terminal phosphate group (γ-PO_4^-) on a phosphoryl donor (usually ATP) resulting in the transfer of the phosphoryl group to the amino acid side chain (Figure 1.10). The reaction is facilitated by the presence of magnesium (Mg^{2+}), which chelates the γ- and β-phosphate groups, lowering the threshold for phosphoryl transfer to the nucleophilic (–OH) group.

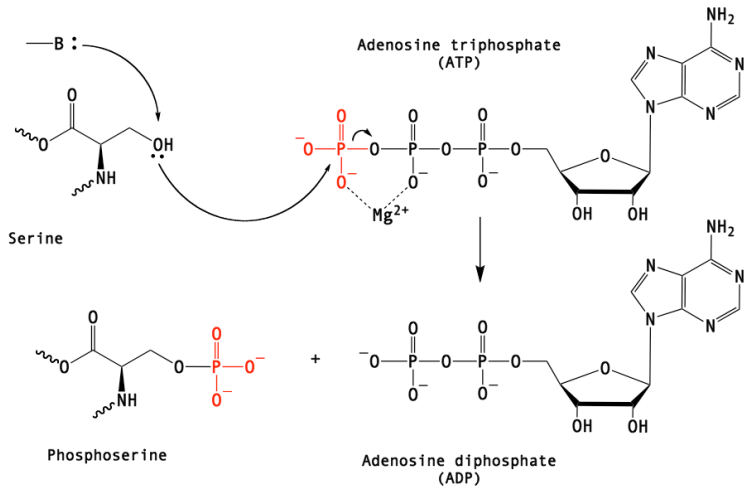


Figure 1.10. Schematic reaction of serine phosphorylation. Kinases catalyse proton transfer from the (–OH) group on serine stimulating the nucleophilic attack of the γ-phosphate group on ATP. As a result, the phosphoryl group is transferred to the serine to form phosphoserine residue and adenosine diphosphate (ADP). The reaction is unidirectional due to the large amount of free energy released when the phosphoanhydride bond in ATP is broken. (—B:) indicates the enzyme base that initiates proton transfer.

Chemical changes caused by phosphorylation can induce conformational changes leading to activation or inactivation of the protein. Alternatively they can

contribute to recruitment of proteins that have structurally conserved domains able to recognize and bind to specific phosphomotifs (phosphotyrosines, phosphoserines and so on). In some cases these phosphoproteins may have a huge ability to recruit other proteins and start a signalling cascade.

Despite the important biological role of PTMs on the function of both E7 and E1A, the impact of phosphorylation on their structure has not been studied. The identification, characterization, and mapping of these modifications through NMR and their effect on the whole protein are very important to understand their functional significance in a biological context.

.

Chapter 2

Methods

2.1 Bioinformatics studies

Nucleic acids (DNA and RNAs) are very important biomolecules for cell function. DNA stores the information for protein synthesis and RNA carries out the instructions encoded in DNA. However, proteins are the primary cellular effectors, responsible for carrying out and mediating most of the biological activities. Thus, the study of proteins is crucial for the understanding of cellular functions.

The collection of physicochemical and structural properties of the biomolecules of interest is generally the first task to accomplish. To help on this, several databases and web software tools have been developed throughout the years and are still in continuous evolution. Nowadays, there are databases and advanced software tools available to provide rich information about DNA sequences, expression profiles, protein sequences and their three dimensional structures or structural ensembles. The tools to predict physicochemical properties of a protein from its primary sequence are very useful. Some of these tools were essential for the development of this work. They are briefly summarized below:

BLAST - The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares a given nucleotide or protein sequence to sequences on databases and calculates the statistical significance of the matches [197].

ExPASy-ProtParam – It computes various physicochemical parameters for a given protein sequence, including isoelectric point (pI), atomic composition, molecular weight (MW) and extinction coefficient (ε) [198].

PROSITE – It is a database of protein families, domains and functional sites, useful to identify functional sites and/or sequence similarities between proteins, often from the same family or from a common ancestor. Thus, for a given protein, PROSITE may indicate the presence of conserved regions and functional sites such as sites for occurrence of PTMs (by identification of specific PTMs sites and/or by similarity analysis) [199].

KinasePhos – It predicts phosphorylation sites within a given protein sequence categorizing these sites by substrate sequences and their corresponding protein kinase classes. Thus, KinasePhos provides not only the phosphorylation site but also the corresponding catalytic protein kinases that could be involved in the predicted phosphorylation [200].

2.1.1 Prediction of disordered proteins/domains

The prediction of the tendency of a protein to adopt stable and well defined 3D structure or to be intrinsically disordered can be achieved through several web servers and bioinformatics tools. Dunker, Uversky and their co-workers developed the first well-tested predictor based on the analysis of protein primary sequences [14,201]. After this, many methods have been designed and different computing techniques have been utilized. Nowadays, several web servers are available for prediction of disorder. They are regularly assessed as part of the Critical Assessment of Techniques for Protein Structure Prediction (CASP), a community-wide experiment for blind protein structure prediction [202]. Besides the developed methods have generally been successful on predicting disorder, their accuracies are variable.

Basically, the methods apply three general prediction strategies: sequence-based approach, machine learning classifiers and meta-prediction. It is important to mention that this is not an absolute classification since some of the methods use more than one of these strategies. The general prediction strategies are briefly described below:

Predictors based on sequence properties – These methods are based on the relative propensity of amino acids to form secondary structures depending on their physicochemical properties. Example of sequence-based methods are IUPred[203], GlobPlot [204], Ucon[205] and FoldIndex[206].

Predictors based on machine learning classifiers – This is a knowledge-based method i.e., it involves extracting knowledge from solved structures existent on databases (such as the PDB) and then use the information achieved to classify new proteins whose structures are unknown. Many machine-learning algorithms for IDPs are available including artificial neural network (ANN) and support vector machine (SVM). Some examples of machine learning systems are the DisPredict, DISOPRED and DISOPRED2 [26,207].

Meta-predictors - They combine the final results reported by series of disorder predictor methods to increase the accuracy of the prediction. Examples of meta-predictors are meta PrDOS [208], PONDR-FIT [101] and DisMeta [209].

In the present work, all predictions were performed using PONDR-FIT, a meta-predictor which combines six predictors (PONDR-VLXT, PONDR-VSL2 [210], PONDR-VL3, FoldIndex, IUPred and TopIDP [17]). The three predictors of the PONDR series use ANN. FoldIndex, IUPred, and TopIDP use sequence-based methods.

2.2 Recombinant protein expression using *Escherichia coli (E. coli)* systems

Genetic engineering has enabled the production of numerous recombinant proteins in large scale by using heterologous systems. It has been essential for several applications such as pharmaceutical industry and drug discovery, food industry, research, etc. The expression of a desired recombinant protein depends on many factors such as the biological activity of both, the recombinant protein itself and the host system used. Today, several host systems are available for protein production, such as bacteria, yeast, fungi, algae, insects, plants, and mammalian cells [211].

The selection of a good host system needs to consider the growth features of cells, eventually PTMs within cells and, of course, the overall expression levels of protein. In this respect, *E. coli* offers many advantages for rapid and economical recombinant protein production such as fast growth kinetics when cells grow in the optimal environmental conditions (media composition and temperature, for example) [212]; easy achievement of high cell density in cultures [212,213] and fast and easy transformation with exogenous DNA [214]. Thanks to these features, *E. coli* is the most used prokaryotic system for production of recombinant proteins [212]. Today, many distinct *E. coli* strains, presenting specific advantages to improve protein expression, are available. The choice of a correct plasmid system and the best *E. coli* strain for each case is critical to obtain the desired protein with a good yield and quality.

2.3 The plasmid systems

In the present work pET systems (plasmid for expression by T7 RNA polymerase) were used as vectors to carry the genes for expression of the desired proteins. pET vectors are derived from the pBR322 plasmid, one of the first used *E. coli* cloning vectors [215]. Nowadays, pET systems are the most widely used systems for cloning and expression of recombinant proteins in *E. coli* [216].

pET systems take advantage of T7 bacteriophage gene *10* (g10-L), which encodes the expression of a ribosome-binding site able to stimulate an efficient translation in *E. coli*. In these vectors the gene of interest, encoding the target protein, is cloned downstream of the T7 promoter and gene *10* leader sequences. The T7 promoter is recognized by the bacteriophage T7 RNA polymerase but not by the host *E. coli* RNA polymerase. In this way, recombinant protein expression is initiated either by transferring the plasmid into an expression host containing integrated copies of the T7 RNA polymerase gene under *lacUV5* control, or by infecting the host cells with λCE6, a bacteriophage that carries the T7 RNA polymerase gene under the control of the λ *p*L and *p*I promoters. In the first case, expression is induced by the addition of isopropyl β-D-1- thiogalactopyranoside (IPTG) to the bacterial culture.

2.3.1 Plasmid of the full-length HPV-16 E7

The plasmid pET20b(+) containing HPV-16 E7 (pET20b_E7) gene was generously provided by Thomas Hey from Scil Protein GmbH, Halle, Germany [217]. The pET20(+) (Novagen) vector includes a sequence for expression of β-lactamases, resulting in ampicillin resistance, an N-terminal *pelB* leader sequence for potential periplasmic localization and an optional C-terminal His-tag (6x Histidine) sequence for further optimization of protein visualization/purification (Figure 2.1).

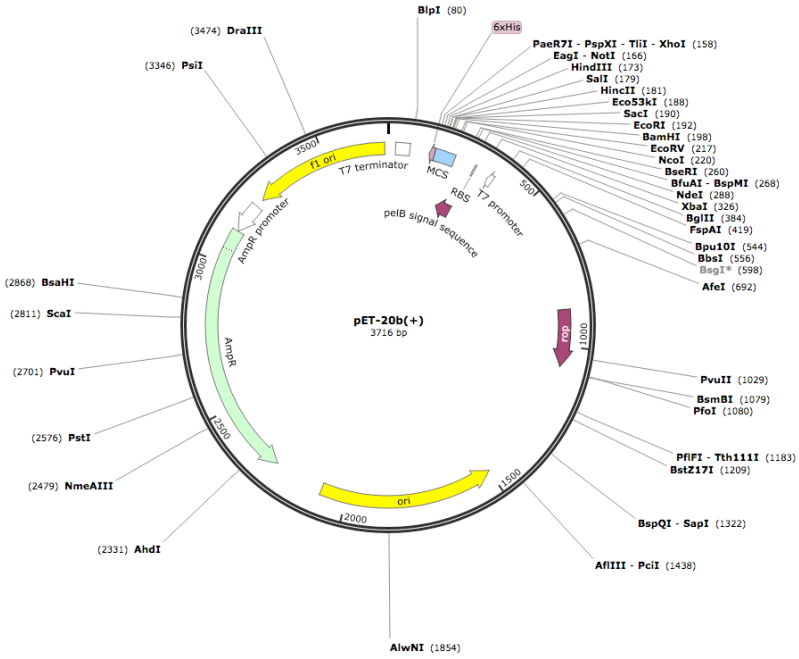


Figure 2.1. Novagen pET-20b (+) plasmid map where HPV-16 E7 construct is inserted.

The original plasmid provided does not generate a native protein (98 residues) but a construct including the C-terminus His-tag. In order to express the wild type form, in a previous work [218] a stop codon (TAG) was inserted right after the sequence of HPV-16 E7 removing the expression of the His-tag.

Protein sequence expressed (HPV-16 E7):

MHGDTPTLHEYMLDLQPETTDLYCYEQLNDSSEEEDEIDGPAGQAEPDRAHY
NIVTFCCKCDSTLRLCVQSTHVDIRTLEDLLMGTLGIVCPICSQKP

DNA sequence (HPV-16 E7):

ATG CAC GGA GAT ACA CCA ACA TTG CAT GAA TAT ATG TTA GAT TTG CAA CCA GAG ACA ACT GAT CTC TAC TGT TAT GAG CAA TTA AAT GAC AGC TCA GAG GAG GAA GAT GAA ATA GAT GGT CCA GCT GGA CAA GCA GAA CCG GAC AGA GCC CAT TAC AAT ATT GTA ACC TTT TGT TGC AAG TGT GAC TCT ACG CTT CGG TTG TGC GTA CAA AGC ACA CAC GTA GAC ATT CGT ACT TTG GAA GAC CTG TTA ATG GGC ACA CTA GGA ATT GTG TGC CCC ATC TGT TCT CAG AAA CCA

2.3.2 Plasmid of E7CR3 and E1ACR3 domains

The plasmids pET21a(+) containing either E7CR3 or E1ACR3 genes (pET21a_ E7 CR3 and pET21a_ E1ACR3, respectively) were synthetized by Eurofins Genomics. pET21a(+) vectors carry optional N-terminal T7-tag sequence and C-terminal sequence. These vectors also include a sequence for expression of β-lactamases, conferring resistance to ampicillin (Figure 2.2).

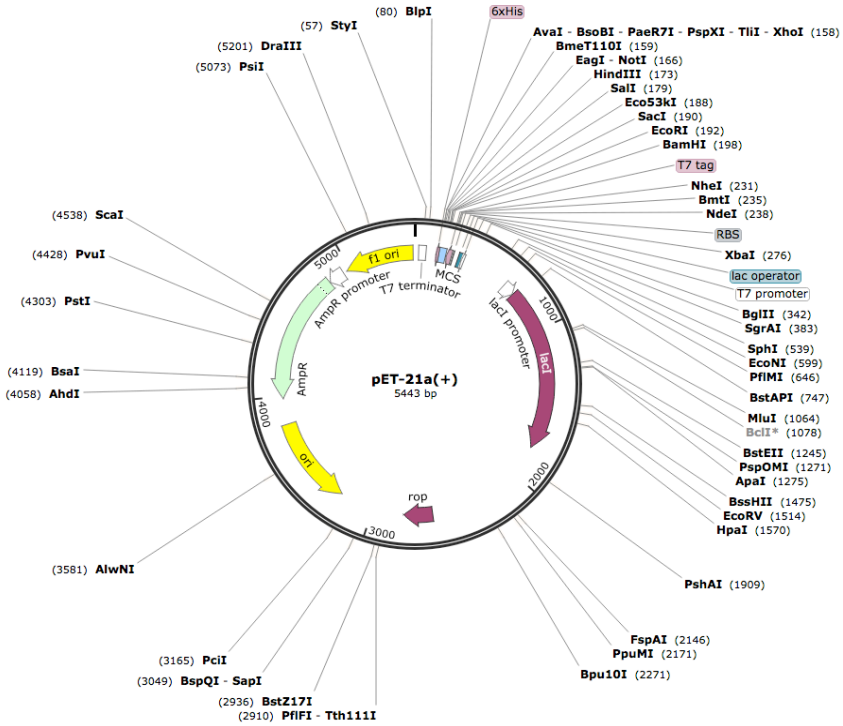


Figure 2.2. Novagen pET-21a(+) plasmid map where E7CR3 or E1ACR3 constructs are inserted.

Protein sequence expressed (E7CR3):

MAEPDRAHYNIVTFCCKCDSTLRLCVQSTHVDIRTLEDLLMGTLGIVCPICSK
P

DNA sequence (E7CR3):

ATG GCG GAA CCA GAT CGT GCC CAC TAT AAC ATC GTC ACG TTT TGC TGC AAA TGC GAT TCG ACC TTA CGC TTG TGT GTG CAA AGT ACC CAT GTA GAC ATT CGC ACA CTG GAG GAT CTC CTG ATG GGT ACT CTG GGC ATT GTT TGT CCG ATC TGT AGC CAG AAA CCG

Protein sequence expressed (E1ACR3):

MDEEGEEFVLDYVEHPGHGCRSCHYHRRNTGDPDIMCSLCYMRTCGMFVYS
PVSE

DNA sequence (E1ACR3):

ATG GAT GAG GAA GGG GAA GAG TTT GTC CTG GAT TAC GTT GAA CAC CCA GGT CAT GGC TGT CGC AGT TGC CAC TAT CAT CGC CGT AAC ACC GGT GAT CCG GAC ATT ATG TGC TCG TTA TGC TAC ATG CGT ACG TGT GGC ATG TTC GTG TAT AGC CCG GTA TCC GAA

2.4 Cell transformation

The first step of protein expression is the cell transformation, i.e. the insertion of the foreign plasmid in the selected host systems. In *E. coli*, efficient cell transformation can be reached through the heat shock method [219]. It consists in incubating the chemically competent *E. coli* cells on ice in the presence of the plasmid of interest for some minutes. After, the mix of bacterial cells and plasmid is placed at 42°C for 45 seconds (heat shock) and then placed back on ice. Then, a rich medium (SOC or LB) is added to the mix and the transformed cells are incubated at 37°C with agitation for some minutes. After, the cells are disposed in a LB-agar plate containing antibiotics for cell selection. The efficient transformed cells present the antibiotic resistance offered by the plasmid inserted and thus, grow on the plate surface, forming colonies.

For further cultures one single colony is chosen to prepare the pre-inoculum. The latter consists of a small volume of LB medium (10-50 mL) supplemented with specific antibiotics (according to plasmid and cell resistances) where the chosen cells grow overnight at 37ºC.

2.5 Expression tests and final cultures

In order to obtain a high yield of expression of the desired protein, different *E. coli* strains were used on small-scale expression tests (2-5 mL of culture). Since the expression tests were realized in small volumes, the cells were added directly to the growth medium (without preparing the pre-inoculum).

The expression tests are very important to choose the most appropriate *E. coli* strains and also to set up the optimal experimental conditions to obtain a soluble protein (preferable) with good yield and quality. Usually, several conditions are evaluated during the expression tests. For example, the culture medium composition, the temperature to grow the culture and to induce protein expression, the concentration of the inducer agent, the optical density of the culture for the induction as well the expression time [220,221].

In the present work, expression tests were performed using BL21(DE3) and BL21(DE3)pLysS, a protease deficient strain. Both *E. coli* strains were transformed with the plasmid for expression of E1ACR3, HPV-16 E7 and E7CR3. Transformed BL21(DE3) and BL21(DE3)pLysS were induced with IPTG at different concentrations (0.25, 0.5 and 1 mM), with an optical density (OD_{600}) of 0.6 and 1.0; after induction, the expression was accomplished at three different temperatures (17, 30 and 37 °C) for three time periods (4h, 8h and overnight).

The evaluation of an expression test includes harvesting and disrupting cells under similar conditions. The solubility of the target protein can be evaluated by SDS polyacrylamide gel (SDS-PAGE). The best choice would be the *E. coli* strain and experimental conditions that could provide a good yield of soluble protein. As cited before, protein solubility depends not only by the toxicity of the protein in the host system but also by the experimental conditions. The composition of the culture medium including its additives is one of the most important factors that can influence

not only the yield but also the solubility of the desired protein. For example, the addition of metal ions such as zinc salts ($ZnCl_2$ or $Zn(CH_3COO)_2$) for expression of a zinc-protein and the amount of carbon and nitrogen sources as well their substitution with the proper reagents for preparation of isotopically enriched ones, can greatly influence the expression levels and/or the formation of inclusion bodies (IBs).

IBs are the result of intracellular accumulation of partially folded and/or insoluble non-native proteins overexpressed within *E. coli* cells. Changes on the growth conditions, such as growth temperature, concentration of inducer, and induction time, can often help in decreasing the formation of IBs. The addition of chemical additives to the culture medium such as sorbitol, arginine, and trehalose could decrease or even suppress the formation of aggregates [222]. However, in some cases, after several expression tests and improvement of conditions, the target protein is still obtained as IBs. Besides IBs are devoid of biological activity, they can be recovered and provide functional and active proteins through elaborated solubilization, refolding and purification procedures.

After the optimal growth conditions are identified, the expression of the target protein takes place on large scale. Usually, the productions of non-labelled proteins without specific requirements may be performed by using a rich medium such as LB to grow cells and induce protein expression. Instead, to express proteins that coordinate metal ions and/or isotopically enriched (^{15}N, ^{13}C and/or ^{2}H) proteins a medium in which the concentration of the various compounds can be controlled is required. The minimal salts (M9) medium* is one of the most common media for these purposes. It can be easily supplemented, for example with metal ion sources. For preparation of ^{15}N-labelled samples, $(NH_4)_2SO_4$ is generally replaced by $(^{15}NH_4)_2SO_4$ in the M9 medium. The same is done for the preparation of ^{13}C-labelled samples where also C_6-Glucose was replaced by $^{13}C_6$-Glucose as source of carbon.

In the present work, all proteins present a zinc-binding site. Due to that, the target proteins were prepared by using M9 media supplemented with 100 μM of

* M9 composition: 48.5 mM Na_2HPO_4, 22.0 mM KH_2PO_4, pH 7,4; 8.5 mM NaCl, 0.2 mM $CaCl_2$, 2.0 mM $MgSO_4$, 100.0 µM $ZnCl_2$, 1 mg/L each of biotin and thiamin, 7.5 mM $(NH_4)_2SO_4$ or $(^{15}NH_4)_2SO_4$ and 11.1 mM C_6-glucose or $^{13}C_6$-glucose)

$ZnCl_2$. In addition, M9 was supplemented with specific antibiotics (See table 2.1). In all the cases, the pre-inoculum was prepared in 50 mL of LB with the specific antibiotic, growing at 37ºC. HPV-16 E7 and E7CR3 cultures were prepared by adding 10-12 mL of pre-inoculum in 1L of M9 medium. When OD_{600} reached 0.7, protein expression was induced by the addition of IPTG. E1ACR3 was instead prepared by the Marley method [223] in order to improve the expression yield. For this, 40-48 mL of pre-inoculum was used to inoculate 4L of LB medium supplemented with ampicillin (0,1 mg/mL). The cultures were allowed to grow at 37ºC in constant agitation (180 rpm) till they reach OD_{600} 0.8. After that the cells were gently harvested (4.500 rpm) for 20 min at 4ºC. Thus, the pellet obtained was re-suspended with M9 medium and allowed for adaptation at 37ºC in constant agitation (180 rpm) for 1hour before the addition of IPTG.

After protein expression, the cell culture should be gently harvested by centrifugation in order to separate the cells (pellet) form the culture medium. The portion that contains the protein may be frozen (-20ºC) before proceeding with the purification steps. Generally this portion is the pellet but in cases of excreted protein, the portion to be stored for purification is the growth medium.

Table 2.1. Parameters for the expression of target proteins:

Protein	Transformed *E. coli* Strain [(a)]/Plasmid	Medium and Antibiotic	IPTG (mM)	Temperature and time of induction
Unlabelled, ^{15}N- and ^{13}C, ^{15}N-labelled E1ACR3	BL21(DE3)/ pET21a_ E1ACR3	M9, Ampicillin [(b)]	0,5	17 ºC; 16 hours
Unlabelled, ^{15}N- and ^{13}C, ^{15}N-labelled HPV-16 E7	BL21(DE3)pLysS/p ET21a_E7	M9, Ampicillin [(b)], Chloramphenicol[(c)]	1	37 ºC; 5 hours
Unlabelled, ^{15}N- and ^{13}C, ^{15}N-labelled E7CR3	BL21(DE3)/ pET21a_ E7CR3	M9, Ampicillin [(b)]	0,25	37 ºC; 5 hours

(a) Stratagene® *E. coli* competent cells; (b) (0,1 mg/mL); (c) (0,34 mg/mL)

2.6 Protein purification strategy

The purification of proteins is an essential step in understanding their structure and function. The use of several and distinct techniques enables to obtain pure protein samples useful for biochemistry and biophysical analysis. Protein purification is not a trivial work. Several characteristics of the protein must be considered before starting with purification trials: isoelectric point (PI), solubility, influence of pH, temperature and salt concentration, the consequences of protein concentration and possible aggregation tendencies, etc. A preliminary bioinformatic analysis as well as consideration based on the behaviour of similar protein systems already described in literature may help in designing the protein purification strategy.

In some cases, special procedures need to be adopted during the purification steps. This is the case of the purification of proteins containing free cysteine residues. The addition of reducing agents such as 2-mercaptoethanol or 1, 4-Dithiothreitol (DTT) in all buffers used during the purification may be necessary. In some cases, the addition of reducing agents is not sufficient to prevent protein oxidation, in particular when the oxidizing agent is the oxygen. In such cases, the purification steps should take place inside an anaerobic chamber.

Taking in consideration a recombinant protein expressed and kept within cells (non-excreted), the first step of protein purification must rely on the disruption of the host cells. First of all it is necessary to re-suspend the host pellet cells obtained after culture harvesting in a lysis buffer. This buffer must be compatible with the later procedures of purification. It usually contains protease inhibitors and possible additives that could guarantee the protein conservation during the initial purifications steps (for example reducing agents).

The disruption of the cells can be performed though several techniques depending on the scale of culture and expression. Usually sonication is largely used for small scale while French press and high-pressure homogenization are used for medium scale and large scale, respectively. In the present work, the expression of proteins was performed in small scale. After cells growth and harvesting, *E. coli* cells were disrupted by sonication. During this procedure the control of temperature is a

critical parameter. However, the use of a series of short pulses incremented with intervals while keeping the cells suspension on ice during all the procedure is a way to avoid the heating problem.

The protein solution obtained after sonication contains membranes, disrupted organelles and, in some cases, IBs (when insoluble proteins are expressed). These components must be separated by centrifugation resulting on a dense pellet and a supernatant. While soluble proteins can be purified directly from the supernatant, insoluble proteins need to be recovered from IBs using denaturant agents such as urea or guanidinium chloride at different concentrations. The last procedure may cause partial or total protein denaturation, thus requiring a refolding step.

Usually, protein mixtures are subjected to a series of separation steps. They can be purified according to their size, charge, solubility and binding affinity. The choice of each technique needs to take into consideration the sample conditions including its impurity level, the amount of protein to be purified as well the yield expected. The chromatography methods are extensively used to reach a high yield of pure protein. Among many existent chromatography techniques, these were mainly used in this work: ion exchange chromatography (IEC), affinity chromatography represented by the immobilized metal ion affinity chromatography (IMAC), and size exclusion chromatography (SEC).

The IEC separates proteins on the basis of their net charge at a given pH. Thus, the protein solution to be loaded into the column needs to follow specific conditions that favours the binding of the target protein to the column, like controlled pH and low salt concentration. To this end, it is very important to known the isoelectric point (pI) of the target protein. Tools such as ExPasy[224] server[*] may be used to calculate the expected pI for a given protein sequence, facilitating the choice of the column resin and buffers to be used on this purification step. At pH values lower than pI, the protein is positively charged and can bind to a negatively charged column resin (cation exchange solid phase) such as carboxymethyl-cellulose (CM-

[*]http://web.expasy.org/protparam/

cellulose). The gradual increase of salt concentration in the eluting buffer (usually NaCl or KCl) makes the cation compete with the positively charged groups on the protein for binding to the column. Proteins that have a low density of net positive charge are eluted first, followed by those having a higher charge density. On the other hand, at pH values higher than the pI, the protein of interest is negatively charged and can bind to a positively charged column resin (the anion exchange solid phase) such as diethylaminoethyl-cellulose (DEAE-cellulose). In a similar way as for cation exchange purification, the gradual increase of salt concentration (usually NaCl or KCl) in the eluting buffer makes the anion compete with the negatively charged groups on the protein for binding to the column. The correct balancing of this amphoteric phenomenon is essential to the success of IEC methods.

Affinity chromatography takes advantage of the high affinity of many proteins for specific chemical groups. One of the most used systems for protein purification is the IMAC [225]. It exploits the affinity of the side chains of specific amino acids (like histidine, cysteine and tryptophan) for metal ions (such as Cu^{2+}, Ni^{2+}, Co^{2+} and Zn^{2+}) attached to the stationary phase (resin) [226]. The interaction of the specific amino acid with metal ions is affected by many variables such as buffer solution composition and pH and temperature during the purification. Moreover, protein purification in IMAC depends on the number and distribution of specific amino acids that can interact with the metal ion as well on protein size [225,227]. Thus, when designing the DNA sequence to express the desired protein, it is important to consider the addition of a sequence for expression of tags in the N or C-terminus as well its cleavage sites (protease recognition site), in order to eliminate the tag after the purification step, if it is necessary.

The first step for protein purification consists in loading the protein solution in the column. Proteins that do not contain a sufficient number of the specific amino acids in an exposed region of the protein do not bind the metal ions immobilized onto the resin and pass through. The purification procedure is based on the elution of the bound protein by a decreasing pH gradient or by an increasing gradient of a competitive agent, such as imidazole, in a buffer. The tag attached to the eluted protein can be cleaved by enzymatic digestion with specific proteases (like TEV,

Thrombin, etc), according to the strategy chosen during the sequence design. After the enzymatic digestion the target protein can be purified from the cleaved tag again with the IMAC. In this case, the cleaved tag remains attached to the column while the target protein is not attached and passes through immediately.

Size exclusion chromatography (SEC), also known as gel-filtration chromatography, is often combined with other techniques like IEC and IMAC for protein purification. It consists in separating the molecules based on their hydrodynamic radius, by passing the mix of proteins through spherical beads (resin) containing pores of a specific size distribution. According to their hydrodynamic properties, proteins may be separated by inclusion or exclusion from the pores of the resin. Molecules larger than the pores size do not enter the pores. They pass between the beads following the solvent flow, being thus eluted first. Instead, smaller molecules diffuse into the pores, being retarded and eluted after larger molecules. The high efficiency on separating molecules by SEC depends of many factors including the column length, pore size, buffer conditions, temperature and pressure of the system, flow rate, and fractionation volume. All these parameters need to be well planned and tested in order to find the best conditions for a high resolution and efficient chromatographic separation.

Since SEC separates proteins based on their radius but not on their molecular weight, the separation of proteins with the same molecular weighs but different shapes is possible. In this case, it is expected that the more globular protein will have a longer retention time than less folded proteins or IDPs. Due to these features, SEC is also used for buffer exchange (desalting) and for the characterization/monitoring of aggregation tendencies and analysis of molecular shape [228]. With the same purpose, SEC may also be combined with other techniques where size/shape based separations are required like dynamic light scattering (DLS), small-angle X-ray scattering (SAXS), etc.

In order to purify HPV-16 E7, the pellet cells obtained after cells harvesting (as described Section 2.5) were re-suspended in 25 mL of degassed buffer A (20 mM TRIS buffer at pH 8.5, 10 mM DTT supplemented with Roche complete protease inhibitors) and bacterial cells were disrupted by sonication (20 kHz sonication

frequency and 40% of amplitude) with cycles of 3s and 10s of delay for 20 min, in anaerobic conditions. Lysed cells were then centrifuged at 165000 *g* for 30 min at 4ºC. The supernatant containing HPV-16 E7 protein was loaded on a HiTrap QFF 10 mL, an ionic exchange column, pre-equilibrated with buffer A. The purification step was followed with buffer A in a gradient with buffer B (buffer A + 0.5 M NaCl). HPV-16 E7 protein was eluted with 0.5 M NaCl (100% buffer B). Fractions containing the protein were mixed, concentrated to 4 mL and injected in a Superdex 75 16/100 size exclusion chromatography column pre-equilibrated with the final degassed buffer (10 mM HEPES pH 7.5, 50 mM KCl, 10 mM DTT and 10 μM $ZnCl_2$). The purification step was followed by using a flow rate of 1 mL/min until the elution of the protein (about 70 mL). After analysis by SDS-PAGE 17%, fractions containing the pure protein were mixed and concentrated to 500 μL for NMR analysis. 50 μL of D_2O was added to each NMR sample for the lock signal. Protein concentration was estimated by refractometer using Lysozyme as standard.

For the purification of E1ACR3 the pellet cells were resuspended in 50 mL equilibration buffer A (20 mM TRIS, 10 mM NaCl pH 8.0, 1 mM DTT and Roche complete mini protease inhibitor). Cells were disrupted by sonication on ice (20 kHz sonication frequency and 80% of amplitude) with cycles of 3s and 10s of delay for 25 min, in anaerobic conditions. Lysate cells were centrifuged at 165000 *g* for 30 min at 4 °C. The supernatant was filtered and loaded on a HiTrap Q FF 5 mL, an ionic exchange column, pre-equilibrated with buffer A (20 mM TRIS, pH 8.0 and 1 mM DTT). The purification step was followed with buffer A in a gradient with buffer B (buffer A + 1 M NaCl). The recombinant E1ACR3 was eluted between 0.25M and 0.35 M NaCl gradient. All the fractions containing the protein were concentrated until 5 mL and loaded on a HiLoad16/60 Superdex 30 prep grade size-exclusion chromatography column pre-equilibrated with 10 mM HEPES pH 7.5, 50 mM KCl, 1 mM DTT and 10 μM $ZnCl_2$. The purification step was followed by using a flow rate of 1 mL/min until the elution of the protein (about 80 mL). After analysis by SDS-PAGE 17%, fractions containing the pure protein were mixed and concentrated to 500 μL for NMR analysis. 50 μL of D_2O was added to each NMR sample for the lock signal. Protein concentration was estimated by UV measurement using the molar

extinction coefficient (5960 $M^{-1}cm^{-1}$) calculated from ExPASy ProtParam tool*. The purification of E7CR3 followed a different protocol, since this protein was expressed in IBs. The detailed protocol of purification and refolding of E7CR3 is described in the next section.

2.7 Refolding Processes

Several approaches have been reported to recover the insoluble target protein in a biologically active form. Despite the protein expressed in IBs has no biological activity, IBs allow a very high level of protein expression with lower degradation inside the cells. Indeed, expression of proteins as IBs may even facilitate the following purification steps, provided of course the refolding procedure is working properly. Since the target protein is delivered in IBs with less impurities, a reduced number of purifications steps are required when compared to proteins expressed in the soluble phase [229].

The refolding process involves the recovery of inclusion bodies from disrupted bacterial cells using low speed centrifugation and their dissolution with a high concentration of chaotropic reagent used to decrease the non-covalent interactions between proteins. Some examples of chaotropic reagents commonly used to solubilize proteins from IBs are urea, guanidinium chloride and ionic detergents, such as *N*-lauroylsarcosine. In addition, reducing agents such as DTT, 2-mercaptoethanol or tris(2-carboxyethyl)phosphine (TCEP) are required to reduce undesirable inter- and/or intramolecular disulfide bonds formation [230]. The solubilisation of proteins from IBs using high concentrations of chaotropic reagents causes loss of tertiary structure leading to the random coil formation and consequently exposure of hydrophobic cores, when present.

Nevertheless, some secondary structures may remain in protein chains [231]. The refolding may be achieved through the slow removal of the chaotropic added reagent, leading the target protein from the completely or partially unfolded state to a folded form. This process may take from a few seconds to several days.

*http://web.expasy.org/protparam/

In order to favour the refolding pathway it is necessary to have under control the most critical parameters that may affect the refolding process:

Protein concentration: the refolding process must be performed with a low concentration (diluted) of the target protein. High protein concentration can result in a massive and irreversible protein aggregation and precipitation.

Salt concentration: although most proteins are less soluble at high salt concentration, in most cases the presence of salt (such as NaCl or KCl) in buffers is required to guarantee protein solubilization. It is very important to work in a safe range of salt concentrations where protein solubility can be maintained without protein precipitation. Usually, 50-100 mM salt is considered a safe range to preform refolding.

pH: protein stability is highly dependent of buffers' pH. The buffers used for the refolding process should be at least one pH unit away from the pI of the target protein in order to guarantee a minimum protein net charge, thus avoiding massive protein precipitation.

Temperature: it influences the speed of folding but also the propensity of aggregation, having thus a dual effect on the refolding process. Each protein is thermodynamically stable in a given buffer system in a limited temperature range. In general, low temperatures are favorable for productive folding (less propensity of aggregation) but they also reduce the folding rates, increasing the time required for refolding. Trials of refolding with different temperatures may be necessary in order to find the best conditions.

Additives: the use of some compounds often helps on the solubility and avoids aggregation during the refolding process. L-arginine (0.4-1 M) may reduce protein aggregation and in a higher concentration range (0.5-2 M) it helps on protein solubilization [232]; also proline [233], polyethylene glycol (PEG) [234] and polyols such as sorbitol [235] and, in particular, glycerol [236] are also used as co-solvents and inhibitors of aggregation.

In the past years many methods have been reported to refold proteins such as dilution, dialysis, chromatography methods reversed micelle systems [237], zeolite absorbing systems [238] as well as the exploitation of natural folding chaperones [239].

Among these approaches, dilution, dialysis and chromatography are the most used methods in the lab routine. SEC is well suited to recover solubilized proteins, since it involves the separation of molecules (depending on their hydrodynamic radius) during buffer exchange, it is able to separate different folding intermediates, reducing protein-protein interaction and thus, prevents aggregation. Indeed, SEC is able to separate soluble aggregates perhaps formed during the previous steps (e.g. solubilization), allowing the obtainment of more homogeneous and pure protein samples.

The E7CR3, expressed in inclusion bodies, was solubilized with 25 mL of 50 mM HEPES, pH 8.5, 50 mM KCl, 2 M urea and 5 mM DTT. After this, the solution obtained was centrifuged at 165000 *g* for 30 min at 4 °C. The supernatant was thus concentrated until 4mL. The purification and refolding of the protein was performed on a Superdex 75 16/100 size exclusion chromatography column pre-equilibrated with the final degassed buffer (50 mM HEPES, 50 mM KCl, 5 mM DTT, 10 μM $ZnCl_2$ at pH 7.5 or 8.5 depending on the analysis to be performed). All the buffers and purification systems were refrigerated at 4°C. The purification was obtained by using a flow rate of 1 mL/min until the elution of the protein (about 70 mL). After analysis by SDS-PAGE 17%, fractions containing the pure protein were mixed and concentrated to 500 μL for NMR analysis. 50 μL of D_2O was added to each NMR sample for the lock signal. Protein concentration was estimated by refractometer using Lysozyme as standard.

During the set-up of the refolding protocol, our tests showed that the addition of zinc ions salts (such as $ZnCl_2$) had no effect on the refolding of the protein. For this reason, no zinc salt was added to the E7CR3's buffers during this procedure.

2.8 Structural and dynamic properties of proteins by NMR techniques

NMR spectroscopy provides a unique tool to investigate the structural and dynamic properties of biomolecules at atomic resolution in solution. NMR signals chemical shifts (CS) provide highly sensitive probes of molecular structure since they are directly related to the local chemical environment of each nucleus itself [240]. The

information contained in CS is useful for the characterization of conformational changes or binding, to figure out regions with secondary structural elements as well as to aid in the refinement of complex structures and in the determination of the tertiary structure of proteins.

In order to obtain all this information it is strictly necessary to correlate the CS observed with the amino acid sequence of the protein under study. Thus, the sequence specific assignment of a protein is the starting point for the investigation of its structural and dynamic properties as well as of binding to other molecules and metal ions. The assignment process begins by the CS assignment of backbone resonances. Once these CS have been obtained, the assignment of side chains can be achieved.

NMR applied to biopolymers generally results in poorly resolved 1D NMR spectra. The addition of one dimension to the 1D NMR spectra spreads overlapping signals increasing the resolution and consequently providing a clearer protein "fingerprint" [241]. However, in most cases the use of multidimensional and multinuclear experiments is necessary to obtain the necessary information or resolution for protein characterization. This is generally the case of proteins larger than 50 amino acids and/or with high flexibility.

Compact structures found in globular proteins create a unique chemical environment for each nuclear spin, resulting in a large CS dispersion. Instead, for IDPs, their highly flexible nature induces extensive conformational averaging, thus causing a reduction of the CS dispersion. This leads to extensive signal overlap, complicating significantly the interpretation of the NMR spectra [242]. The overlap problem frequently found in 2D NMR spectra can be significantly reduced by the addition of more indirect dimensions, generating for example three and four-dimensional spectra. These strategies rely on coherence transfer between $^{1}H^{N}$,^{15}N and ^{13}C nuclei along the backbone chain.

2.9 Sequence specific assignment and structure determination

Sequence specific assignment strategies are well established for globular proteins where combinations of triple-resonance experiments linking $^{1}H^{N}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N CS provide the necessary information for connecting neighbouring residues. For example, one can combine 3D HNCA and 3D CBCA(CO)NH experiments. While 3D HNCA provides intra- and inter-residue correlations between $^{1}H^{N}$, ^{15}N, and $^{13}C^{\alpha}$ resonances, 3D CBCA(CO)NH gives only inter-residue correlations between the $^{1}H^{N}$ and ^{15}N resonances of one residue and the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ resonance of the preceding residue. Hence, each $^{13}C^{\alpha}$ resonance is linked to both the $^{1}H^{N}$, ^{15}N resonances of intra-residue and preceding amino acids and, in addition, $^{13}C_{\beta}$ resonance information is achieved [243,244]. Ambiguities caused by possible CS degeneracy can be solved with additional experiments that provide alternative correlations such as the 3D HNCO and 3D HN(CA)CO that correlate the $^{1}H^{N}$ and ^{15}N resonances with intra-residue and the sequential $^{13}C'$ resonances. After getting information about the backbone, the side chains assignment can be accomplished with 3D HCCH-TOCSY and ^{1}H-^{15}N TOCSY-HSQC experiments [243]. This is a very important step to verify sequential assignment and to get information for further structure determination.

Triple-resonance experiments acquired on structured proteins that give $^{1}H^{N}$, ^{15}N resolved spectra work well but when they are applied to IDPs they are often doomed to failure due to two reasons: (i) the significant chemical exchange of exchangeable protons with the solvent protons which reduces $^{1}H^{N}$ signal intensities (amide signal broadening resulting in low signal-to-noise ratios), (ii) the extensive signals overlap. In order to address these problems, specific NMR experiments were designed. Increasing dimensionality of the experiments is a solution to increase resolution. The acquisition of 5D - 7D NMR experiments is possible for IDPs thanks to their reduced transverse relaxation rates (compared to structured proteins of similar size).

One approach adopted to reduce the influence of proton chemical exchange on line broadening and increase resolution when working with IDPs, is to use ^{13}C detection [245]. These experiments allow also efficient assignment of proline residues

which are very abundant in IDPs and thus are considered the most disorder-promoting residues [246].

Once chemical shift assignment is obtained, CS can be used to measure backbone dihedral angles Φ and Ψ [247,248] for structured proteins and to estimate secondary structure propensity (SSP) [249] in case of disordered proteins.

The SSP score is based on the deviation of the CS of various resonances of the protein under investigation from their expected random coil values [250]. In particular $^{13}C^{\alpha}$, $^{1}H^{\alpha}$ and $^{13}C^{\beta}$ nuclei are considered good probes for secondary structure since their dispersion is much greater in folded proteins than in IDPs [251]. The SSP score ranges from --1 to 1, representing the tendency of a protein fragment for a specific secondary structure element. Positive values indicate the formation of α-helix structure and negative values indicate the formation of β-strand or extended structure (Figure 2.3). SSP can be calculated in various ways through easily accessible online tools such as through the ncSPC* (neighbour-corrected Structural Propensity Calculator) [249] .

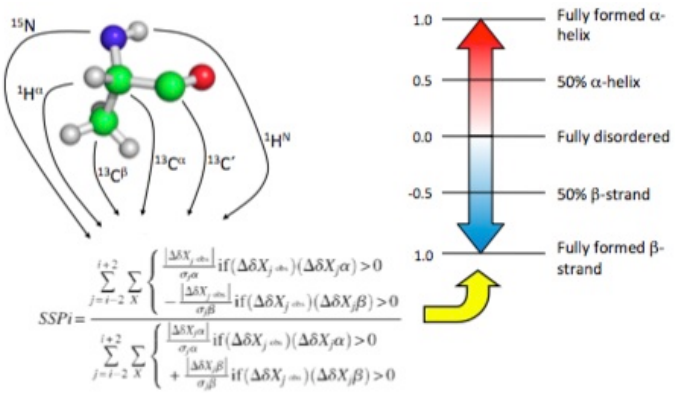


Figure 2.3. SSP integrates different backbone and side-chain nuclear chemical shifts to generate a single estimate of fractional secondary structure content. Adapted from Marsh *et al.*, 2006 [252].

Protein structure determination by NMR requires the measurement of parameters sensitive to distances between nuclei and molecular conformation. At

* http://www.protein-nmr.org

present, the nuclear Overhauser effect (NOE) and the residual dipole–dipole couplings (RDCs) measurements are the most commonly utilized parameters for protein structure determination.

2.10 Nuclear Overhauser effect

The main source of geometric information used in NMR structure determination of biomolecules lies in short inter-proton distance restraints derived from *Nuclear Overhauser effect* (NOE) measurements. NOE results from cross-relaxation due to the dipole-dipole interaction between two nearby spins [253]. Since the NOE effect is proportional to the inverse sixth power of the distance between the two protons**,** it is measurable only for distances up to about 5Å i.e. the NOE is useful to give both short range and local information as well as distances between two protons belonging to different amino acids which may be far away from each other in the primary sequence of the protein but that are close in space.

In two- or higher-dimensional heteronuclear resolved ^{1}H-^{1}H-NOESY spectra the NOE is manifested by cross-peaks between pairs of protons at a short distances one from the other. The evaluation of cross-peak intensity provides an estimate of the inter-proton distances. Thus, besides giving information on distance constraints, NOESY cross-peaks are very useful to verify or complete the sequence specific assignment and eliminate possible ambiguities [254,255]. Indeed, sequentially neighbouring amino acids are also close in space and in general provide inter-residue correlations. In addition, short proton–proton correlations are specific for each type secondary structure. Thus, their identification helps to define local structural features such as the presence of α-helices and β-strand. α-helices are characterized by a stretch of strong or medium $H^N(i)$ - $H^N(i\pm1)$, weaker $H^N(i)$ - $H^N(i\pm2)$ and $H^N(i)$ - $H^N(i\pm3)$ and $H^\alpha(i)$ - $H^N(i\pm1)$ while β-strands are characterized by strong $H^\alpha(i)$ - $H^N(i\pm1)$ and the absence of other short range NOEs involving the H^N and H^α protons [256].

The NOESY cross-peak intensity may be affected by several factors such as protein or domain mobility, solvent exchange and spin diffusion, making invalid the linear correlation between intensity and r^{-6}. Nevertheless, if NOESY spectra have

been acquired in order to avoid severe spin-diffusion effects (by using shorter mixing times) the intensities can be converted into the corresponding $^{1}H–^{1}H$ distances. However, additional problems may arise from spectral artefacts and noise as well from the absence of some expected signals due to the fast relaxation. Due to these inevitable shortcomings interactive procedures are necessary for 3D protein structure determinations.

2.11 Residual Dipolar Coupling

Dipolar couplings are sensitive to the chemical environment and molecular motions. They are a useful probe of molecular structure and dynamics associated with biological processes such as ligand binding, conformational exchange and protein-protein interactions.

Residual dipolar couplings (RDCs) are complementary to NOEs, being used as restraints in molecular dynamics calculations. However, RDCs report on the orientation of dipolar coupling vectors within a common reference frame. In solution NMR and under isotropic conditions, inter-nuclear dipolar couplings average to zero as a result of the effects of Brownian motion. In order to reach a measurable dipolar coupling in solution, protein molecules must be weakly aligned relative to the field, creating an anisotropic condition.

The partial orientation of molecules to enable the determination of residual dipolar couplings is basically created by two ways: one is the direct observation of biomolecules that have a large magnetic susceptibility anisotropy such as metalloproteins with paramagnetic centres or diamagnetic systems, e.g. DNA, where its interaction with the magnetic field produces a weak alignment [257,258]. In the same context, proteins can be also linked to a metal-chelating tag [259,260]. Another way is to make use of an orienting medium that contains small amounts of orienting agents (substances with a large magnetic anisotropy) that could weakly (almost negligibly) interact with the target protein.

2.11.1 Alignment media

RDCs resulting from external alignment media are usually extracted from measurements of observed scalar couplings in the presence and absence of an alignment medium. Many different alignment media have been employed to generate the anisotropic molecular orientation including stretched acrylamide gel (SAG) [261,262], liquid crystalline phases based on virus particles (in particular filamentous phage) [263,264], phospholipids (bicelles) and surfactants, in particular poly(ethylene glycol)-(PEG-) based bilayers [265,266].

The choice of the alignment medium is a not obvious and it is frequently empirically determined. Some criteria must be evaluated such as (i) the chemical compatibility of the system considering buffer conditions (pH and ionic strength) (ii); the stability (the orienting medium must remain stable and oriented, not leading to phase separation) (iii) the possibility of strong interaction between the systems, which must be avoided. Despite none of the alignment media is universally applicable, liquid crystalline media based on bicelles and PEG-based bilayers have been widely used for protein solutions in NMR due to their high versatility.

Bicelles are disk-shaped particles made from a mixture of long- and short-chain phospholipids (usually the long-chain is 1,2-dimyristoyl-*sn*-glycero-3-phosphocholine, DMPC, and the short-chain is the detergent 1,2-dihexanoyl-*sn*-glycero-3-phosphocholine, DHPC) [267,265]. The formation of a liquid crystalline phase with DMPC/DHPC mixtures is very complex and depends on many variables including the DMPC:DHPC ratio and their absolute concentration as well as temperature, salt concentration and buffer pH [268–271].

In solution-state NMR the ratio of DMPC:DHPC is typically 3:1 and the absolute concentration in solution is usually 3-5% (w/v). These conditions allow obtaining small and fast-tumbling bicelles, with a minimal impact on line widths. The use of DMPC/DHPC bicelles has some limitations. The phospholipids are only stable in a narrow pH range (around 6.5). Outside this range, hydrolysis can take place in some days or weeks. Clearly it can be a trouble when working with proteins not stable at this conditions. Indeed, bicelles adopt liquid crystalline phase behaviour, with and

optimal alignment, only over a narrow range of temperatures and ionic strength [272]. These ranges extremely depend on the DMPC:DHPC ratio and on their absolute concentration in solution. Bicelles can be constituted by other phospholipids, for example the 1,2-di-O-dodecyl-*sn*-glycero-3-phosphocholine (DIOPC) and 3-([3-chloramidopropyl]dimethylammonio)-2-hydroxyl-1-propanesulfonate (CHAPSO). Although they are more stable at low pH, they still do not form stable crystalline phase at neutral pH [273].

Upon these conditions, the PEG-based bilayers are useful alternatives for the use of bicelles in solution-state NMR. The main advantage relative to bicelles is that they are characterized by high stability. The PEG-based systems are uncharged and thus insensitive to pH and little sensitive to salt concentrations. Thus, they may remain as liquid crystalline phase for long times, even years. Indeed, PEG-based systems have little or no interaction with macromolecules [266].

PEG-based bilayers are formed by the mixture of PEG and *n*-alkyl alcohol (usually *n*-octanol or *n*-hexanol). The different PEGs are denoted as C*m*E*n*, where *m* is the number of carbons in the linear alkyl chain and *n* is the number of the glycol units in the poly(ethylene glycol) moiety (e.g. C12E6, C12E5 and C8E5). Molar ratios of C*m*E*n* to *n*-alkyl alcohol are generally in the 0.64-0.96 range and the weight percentages of C*m*E*n* are between 3 and 11% [274]. These bilayers have shown a large temperature range (below 0 °C up to almost 40 °C) as demonstrated in Figure 2.3. In fact, the temperature range depends on the *n*-alkyl alcohol concentration, which lowers the temperature range of stability for the liquid crystalline phase.

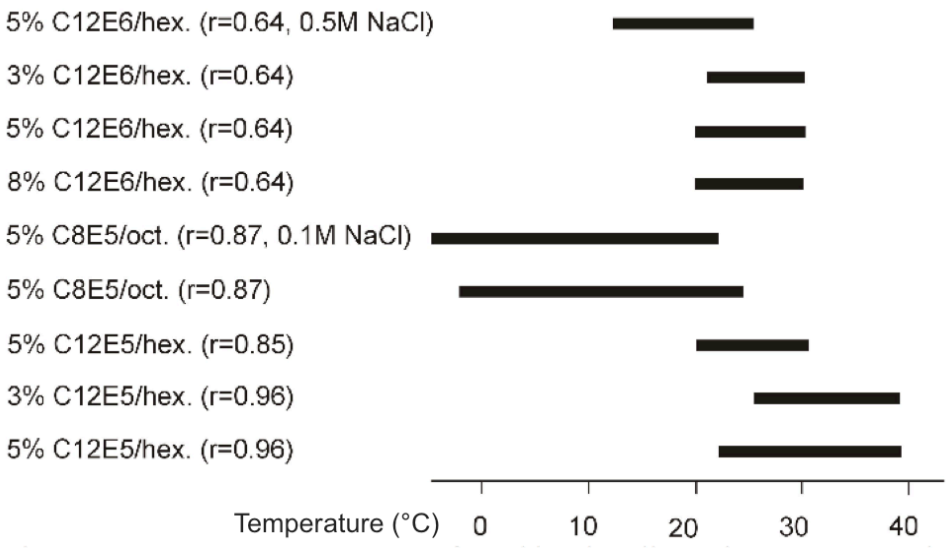


Figure 2.3. Temperature range of stable liquid crystalline phases based on C*m*E*n*/alcohol/H_2O. The systems are identified by the surfactant/H_2O ratio in weight percentages (wt %) taking into account the presence of D_2O, but not the alcohol. The r-value denotes the molar ratio surfactant to alcohol. Adapted from Rückert and Otting, 2000 [266].

In NMR spectra, RDCs can be easily determined by measuring the additional contribution to splitting deriving from the scalar *J* couplings (Figure 2.4). Thus, the magnitude of the dipolar coupling between two nuclei, A and B (D_{AB}), can be easily measured by taking the difference of the splitting under anisotropic conditions ($J + D_{AB}$) and under the isotropic conditions isotropic state (J) [275].

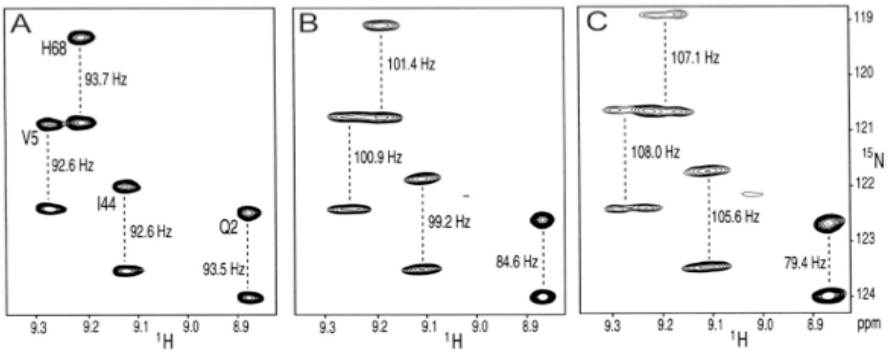


Figure 2.4. Small regions of ^1H-^{15}N HSQC spectra of Ubiquitin protein collected without $^1J_{NH}$ decoupling in the following conditions: A) Isotropic; B) In crystalline liquid medium 4.5% (w/v) of bicelles DMPC/DHPC and CTAB on the proportion 30:10:1, respectively; C) In 8% (w/v) of bicelles DMPC/DHPC and CTAB on the proportion 30:10:1, respectively. The ^1H-^{15}N HSQC scalar coupling are indicated in Hz. Adapted from Bax, 2003 [276].

In the present work, the RDCs measurements of HPV-16 E7 were performed by creating a crystal medium using C8E5 (pentaethylene glycol monooctyl ether) and *n*-octanol. These components were added to the NMR tube in order to have a final solution of 5% w/v (C8E5/buffer) and a molar ration (r) of C8E5 to *n*-octanol of 0.87. The experiments were performed using variants of 2D ^{1}H-^{15}N HSQCs in which the information about the coupling was encoded in the indirect ^{15}N dimension of the spectra through the IPAP approach. Two subspectra were acquired, one collecting the in-phase (IP) component and one the anti-phase (AP) component. Linear combination of the two subspectra allowed to separate the two multiplet components and to measure RDCs for the majority of the residues in the N-terminal part of E7.

2.12 Dynamic studies

The intrinsic dynamic properties of proteins play a key role in their functions and thus they represent an important feature to be investigated. NMR can provide dynamic information in several timescales (ps-ns) by measuring relaxation rates, which depend on molecular motions [277]. NMR relaxation depends on the physical bond vibrations, side chain rotamer interconversion, the presence of random coil, loop motions and backbone torsion angle rotation [278].

^{15}N nuclear spin relaxation is used to probe protein backbone dynamics. ^{15}N R_1 ($1/T_1$, the longitudinal relaxation time constant), ^{15}N R_2 ($1/T_2$, the transverse relaxation time) and the heteronuclear nuclear Overhauser effect *(^{1}H-^{15}N hnNOE)* are generally measured [279].

^{15}N R_1 and ^{15}N R_2 data utilize a series of spectra that quantify the time-dependent decay of magnetization due to spin-lattice relaxation rates (R_1) occurring along the z axis and via spin-spin relaxation rate (R_2) occurring in the x,y plane. The *hnNOE* results from the magnetization transfer via dipolar coupling between ^{1}H and the directly attached ^{15}N (relaxation properties of backbone). It is obtained by comparing signal intensities in the presence and absence of ^{1}H radiofrequency

irradiation [280]. These parameters provide information on local structure, rigidity and chemical exchange.

The ^{15}N R_2 and ^{15}N R_1 ratios (R_2/R_1) are useful to estimate the effective rotational correlation time (τ_c) in proteins with molecular weights up to about 30 kDa [281]. In folded proteins, τ_c can be correlated with a good approximation to the molecular weight and thus be used as an indicative measure of possible aggregation.

The ^{15}N relaxation experiments to determine R_1, R_2 and ^{1}H-^{15}N NOEs of E7CR3 were acquired on a Bruker AVANCE 500 MHz spectrometer (11.8 T) at 298 K. E7CR3 sample was 300 μM concentrated in 50 mM HEPES (pH 8,5), 50 mM KCl, 5 mM DTT, 10 μM $ZnCl_2$.

The ^{15}N relaxation experiments to determine R_1, R_2 and ^{1}H-^{15}N NOEs of E1ACR3 were acquired on a Bruker 700 MHz spectrometer at 288 K on a 140 μM E1ACR3 in 10 mM HEPES (pH 7,5), 50 mM KCl, 1 mM DTT, 10 μM $ZnCl_2$

2.13 Computational approaches to describe the structural and dynamic properties of IDPs

The lack of a stable 3D structure, a characteristic feature of highly flexible and heterogeneous proteins, prevents the use of well-established computational approaches to study IDPs. In principle, molecular dynamics simulations can generate ensembles of structures once accurate energy functions and efficient sampling methods are provided. In practice, however, often the energy functions are known only approximately and the sampling can be carried out only in a limited manner [282]. Indeed, molecular dynamics suffers of a number of problems resulting from an incomplete conformational sampling [283]. The determination of meaningful descriptions of the structural and dynamic properties of IDPs is still a very challenging task. Two different methods were used in this work for the characterization of E1A and HPV-16 E7. In the case of E1A its large size prevents the use of computationally intensive methods; NMR data were thus used as input for

flexible-meccano calculations [284], which nicely allowed us to visualize the information content of experimental data.

In case of E7, the more limited size of the protein allowed us to exploit replica-averaged metadynamics (RAM), a computationally intensive approach, for the determination of the conformational energy landscape of the protein. This method enables to increase the quality of the energy functions and of the extent of the sampling. The method is based on the incorporation of experimental data as replica-averaged structural restraints in molecular dynamics simulations and exploits the meta-dynamics framework to enhance the sampling [282]. This approach in combination with NMR CS and RDCs was used for the characterization of the structural and dynamic features of HPV-16 E7 thanks to a collaboration with the group of Michele Vendruscolo.

.

2.14 Phosphorylation of HPV-16 E7 followed by time-resolved experiments

As mentioned in Section 1.3, NMR is a powerful technique to identify and follow PTMs, allowing the investigation of kinetics and structural changes without further requirements for selective enrichment or purification procedures, as would be required for MS analyses.

The biological consequences of the phosphorylation of Ser31 and Ser32 of E7 are well known. It is an important mechanism of E7 on improving binding with key partners and increasing its oncogenic potential. This phosphorylation reaction requires the presence of Mg^{2+} to form the active Mg^{2+}-ATP complexes. Indeed, Mg^{2+} is required for CKII activity [285,286]. In previous studies performed by our group we have observed that high concentrations of $MgCl_2$ causes loss of signals in NMR spectra of the CR3 domain. Thus, in order to establish the optimum Mg^{2+} concentrations to perform HPV-16 E7 phosphorylation, an $MgCl_2$ titration of HPV-16 E7 and phosphorylation trials were performed with ^{15}N-labeled samples. The titration was followed by acquiring a series of SOFAST-HMQC collected at 298 K on Bruker

Avance 900 MHz instrument when 108 μM E7 were mixed with different $MgCl_2$ concentrations: 0.1; 0.5, 1; 5; 10; 15; 20 and 50 mM.

In vitro HPV-16 E7 phosphorylation was performed with ^{15}N-labelled samples with different concentrations of ATP (15, 25, 35, 350 and 800 μM) from Sigma-Aldrich®; and also different concentration of recombinant human CKII (2500 and 5000 UI) from New England Biolabs®, in presence of 1 mM $MgCl_2$. The reagents were added directly in the NMR tube and monitored by NMR spectroscopy. The reaction was followed with uninterrupted cycles of 1D ^1H [36], ^1H-^{15}N SOFAST-HMQC [37;38] and ^1H-^{15}N HSQC [27-29;39] NMR experiments acquired over the time until the end of the reaction (disappearance of signals of unphosphorylated Ser31 and Ser32). All NMR experiments were acquired at 303 K on a 16.7 T Bruker AVANCE 700 equipped with a cryogenically cooled triple resonance probehead. ^1H-^{15}N SOFAST-HMQC experiments were recorded with 16 transients and 1,204 (^1H) x 128 (^{15}N) complex points. ^1H-^{15}N HSQC were recorded with 4 transients and 1,536 (^1H) x 256 (^{15}N) complex points. Each experiment required 20 min of acquisition time. NMR data were processed with TopSpin 2.0 and analysed with the program ccpNMR[34].

2.15 Sample preparation and their analysis/assays

The expressions and all steps of purification of each protein were designed considering the analyses to be performed. In particular, the last step of purification (usually SEC) was optimized to obtain the target protein in the optimal buffer conditions. Further modifications such as pH adjustments or buffer exchanges were performed under anaerobic conditions by using Millipore (Amicon®) ultra-centrifugal filters to wash the protein with the desired buffer until reaching the desired conditions. The general buffer content for each protein is summarized below:

Sequence specific assignment

E7CR3: 250 μM E7CR3 in 50 mM HEPES (pH 8,5), 50 mM KCl, 5 mM DTT, 10 μM $ZnCl_2$.

E1ACR3: 140 μM E1ACR3 in 10 mM HEPES (pH 7,5), 50 mM KCl, 1 mM DTT, 10 μM $ZnCl_2$.

Dynamic studies

HPV-16 E7: 170 μM HPV-16 E7 10 mM TRIS (pH 8,5), 50 mM KCl, 10 mM DTT, 10 μM $ZnCl_2$.

E7CR3: 300 μM E7CR3 in 50 mM HEPES (pH 8,5), 50 mM KCl, 5 mM DTT, 10 μM $ZnCl_2$.

E1ACR3: 140 μM E1ACR3 in 10 mM HEPES (pH 7,5), 50 mM KCl, 1 mM DTT, 10 μM $ZnCl_2$.

RDC analysis

HPV-16 E7: 150 μM HPV-16 E7 20 mM HEPES (pH 7,5), 50 mM KCl, 10 mM DTT, 10 μM $ZnCl_2$.

$MgCl_2$ titration and phosphorylation assays

HPV-16 E7: 125 – 340 μM HPV-16 E7 20 mM HEPES (pH 7,5), 50 mM KCl, 10 mM DTT, 10 μM $ZnCl_2$.

Chapter 3

Results

The biological importance of HPV-16 E7 and HAdV-E1A proteins (E1A289 and E1A243) and the lack of high-resolution structural and dynamic data on these two important proteins motivated this research project. Besides being involved in different pathologies, HPV-16 E7 and HAdV-E1A proteins share a lot of similarities, in particular interactions with the same key partners and structural elements.

The complete structural and dynamic characterization of both proteins represents an important step for the understanding of their properties at the molecular level. Some of the highly flexible domains of HPV-16 E7 and HAdV-E1A proteins were previously characterized by NMR [56,287] However, the CR3 domains of both proteins revealed to be a challenging target for high resolution characterization. In order to investigate CR3 in both cases, the isolated domains (E7CR3 and E1ACR3) were expressed, purified and analysed by NMR.

The complete sequence specific assignment of E1ACR3 was obtained and thus it was possible to access dynamic information on the specific residues as well as to perform SSP calculations. The results indicate that CR3, in the short construct, is a relatively flexible domain composed by two helices of less than ten residues each, each of them hosting one of the two CXXC zinc-binding motifs. The complete high-resolution characterization of E1A is described in detail in Section 3.1 ("Structural and dynamic characterization of the molecular hub E1A from human adenovirus", published in Chemistry, 2016).

In case of HPV-16 E7, the strategy of expressing the isolated E7CR3 allowed achieving the sequence specific assignment of most parts of this domain, and thus to calculate the SSP. This analysis shows the presence of two α-helixes in regions from

residue 69 to 82 and 92 to 96 in agreement with the structural investigations on the other HPV serotypes [54,77]. ^{15}N relaxation data were also collected and are reported in the Appendix. Since the structural properties of the E7CR3 showed to be quite similar to the CR3 domain in the full-length E7, the assignment obtained in the isolated domain could be transferred to the full-length E7. This represents the assignment of about 90% of the amino acids of HPV-16 E7 ($^{1}H^{N}$ and ^{15}N, $^{13}C'$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ resonances). In order to access complementary structural information of HPV-16 E7, RDCs were measured. For this analysis (C8E5) and *n*-octanol were used to induce weak orientation of the protein. The resulting RDCs together with CS information were used as structural restrains on replica-average metadynamics (RAM). This part of the work was possible thanks to the collaboration with Prof. Michele Vendruscolo from University of Cambridge.

Through RAM methodology it was possible to reconstruct the free energy landscape of the protein and to obtain a conformational ensemble that describes the dynamics of its domains. Our results show that the N-terminus of HPV-16 E7 is characterized by high disorder content while the CR3 domain is more compact. On this domain our results also indicate the presence of two α-helices and one β-sheet and also small linkers that could behave as coiled regions, providing also certain flexibility to this folded domain. The high quality of the 3D models obtained can be useful to understand and design potential anti-HPV drugs to target HPV-16 E7. This work is described in detail in Section 3.2 (The structure and dynamics of HPV-16 E7 described by NMR-guided metadynamics draft in preparation).

The resulting assignment of HPV-16 E7 allowed also to investigate the impact of one of the most important PTMs on E7, the CKII-dependent phosphorylation occurring in Ser31 and Ser32. As mentioned in Section 1.3.2, the phosphorylation of E7 may enhance the affinity of E7 for important key partners, in particular pRb [48] and TBP [65,69]. The phosphorylation of Ser31 and Ser32 monitored through time-resolved NMR experiments showed that the phosphorylation of Ser31 and Ser32 results in minor changes to the local residues. During and after the reaction no conformational changes were observed. The ensemble of these results suggests that the enhancement of affinity of HPV-16 E7 to its partners is due more to the negative charge provided

by the phopshoryl group added by phosphorylation. Sample checks performed after the phosphorylation of Ser31 and Ser32 revealed that the protein remains stable enough time to perform further experiments such as interaction studies. The work HPV-16 E7 phosphorylation is described in Section 3.3 ("Toward the real-time monitoring of HPV-16 E7 phosphorylation events", submitted for publication)

Section 3.1

Structural and dynamic characterization of the molecular hub E1A from human adenovirus

Tomáš Hošek[a], Eduardo O. Calcada[a], Marcela Oliveira Nogueira[a], Michele Salvi, Talita Duarte Pagani, Isabella C. Felli[a] , Roberta Pierattelli[a]

[a] *Magnetic Resonance Center (CERM) and Department of Chemistry "Ugo Schiff", University of Florence, Italy*

Published in Chemistry [288]

3.1. STRUCTURAL AND DYNAMIC CHARACTERIZATION OF THE MOLECULAR HUB E1A FROM HUMAN ADENOVIRUS

ChemPubSoc Europe

DOI: 10.1002/chem.201602510

CHEMISTRY A European Journal Communication

NMR Spectroscopy

Structural and Dynamic Characterization of the Molecular Hub Early Region 1A (E1A) from Human Adenovirus

Tomáš Hošek[+], Eduardo O. Calçada[+], Marcela Oliveira Nogueira, Michele Salvi, Talita Duarte Pagani, Isabella C. Felli,* and Roberta Pierattelli*[a]

Abstract: The small-DNA human adenovirus encodes one of the most versatile molecular hubs, the E1A protein. This protein is essential for productive viral infection in human cells and a vast amount of biologically relevant data are available on its interactions with host proteins. Up to now, however, no high-resolution structural and dynamic information on E1A is available despite its important biological role. Among the different spliced variants of E1A, two are expressed at high level in the early stage of infection. These are 243 and 289 residues isoforms. Herein, we present their NMR characterization, showing that they are both highly disordered, but also demonstrate a certain heterogeneous behavior in terms of structural and dynamic properties. Furthermore, we present the characterization of the isolated domain of the longer variant, known as CR3. This study opens the way to understanding at the molecular level how E1A functions.

Adenovirus early region 1A (E1A) is the first protein to be expressed by the human adenovirus (HAdV) following viral infection and can generate diverse responses in host cells by binding to a large number of cellular proteins.[1–3] Originally identified as an oncogene in rodents,[4] it was later shown to act as a tumour suppressor in specific human cells[5] posing fundamental questions on the molecular origins of this apparently controversial behaviour and demanding further investigation on the molecular determinants of its function.[6] The E1A gene actually codes for two major proteins that are expressed from two alternatively spliced mRNA species (13S and 12S). The 13SmRNA codes for a 289 amino acids protein (E1A289) and the 12SmRNA codes for a 243 amino acids protein (E1A243).[7] The two proteins differ by the presence of a 46 amino acids module that plays a crucial role in viral replication through transcriptional activation of other viral early genes.[8–11] Both E1A289 and E1A243 can immortalize primary cells and can transform them in co-operation with other viral and cellular oncogenes[12] but detailed functional studies on the longer constructs are very limited.[6]

Amino acid sequence alignment of E1A289 from different serotypes reveals four main conserved regions (CR1-CR4),[2,13] as schematically indicated in Figure 1, which are separated by

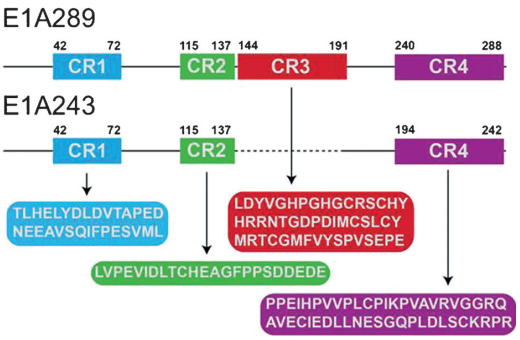


Figure 1. E1A289 and E1A243 amino acid sequences, and their classification in conserved regions (CR).

less conserved regions. E1A243 misses the entire CR3. The latter contains two highly conserved CXXC motifs that are responsible for Zn^{II} binding.[14] The primary sequence of E1A has been used to predict through bioinformatic tools the tendency of the different parts of the polypeptide chain to adopt more ordered or disordered conformations.[15,16] These clearly indicate that E1A is not expected to adopt a stable 3D structure.[2,13,17] E1A is involved in protein–protein interactions with a multitude of cellular factors through short linear motifs (SLiMs) modulating signalling outcomes.[18,19] In this way E1A functions as a molecular hub that can bind promiscuously to numerous cellular proteins and organize them into higher-order complexes that efficiently disrupt regulatory networks and hijack cell regulation.[2,20]

Despite all the information available on E1A and on its documented interactions with key players in host cells, there is no high-resolution structural and dynamic information available on the whole protein.[2] The 3D structure could not be resolved by X-ray crystallography as the protein does not crystalize, confirming the presence of conformational heterogeneity. The investigation through NMR spectroscopy has so far focused on

[a] Dr. T. Hošek,[+] Dr. E. O. Calçada,[+] M. O. Nogueira, M. Salvi, Dr. T. D. Pagani, Prof. I. C. Felli, Prof. R. Pierattelli
CERM and Department of Chemistry "Ugo Schiff"
University of Florence, via Luigi Sacconi 6
50019 Sesto Fiorentino, (Italy)
E-mail: felli@cerm.unifi.it
pierattelli@cerm.unifi.it

[+] Both authors contributed equally to this work

Supporting information and the ORCID identification number(s) for the author(s) of this article can be found under http://dx.doi.org/10.1002/chem.201602510.

ChemPubSoc Europe

CHEMISTRY A European Journal Communication

protein fragments involved in specific protein–protein interactions,[21] since the entire protein is difficult to handle and prone to oligomerization. Notwithstanding these limitations, we decided to undertake the NMR investigation of E1A243 and E1A289 exploiting the recently developed NMR strategy combining ^{1}H and ^{13}C detection experiments.[22–26]

The 2D ^{1}H–^{15}N correlation NMR experiments acquired on E1A289 and E1A243 (Figure 2A,B) confirm that both isoforms of E1A are highly disordered and flexible, as inferred from the small chemical shift dispersion of the NMR signals.[27] A second observation is that the spectra of the two forms are remarkably

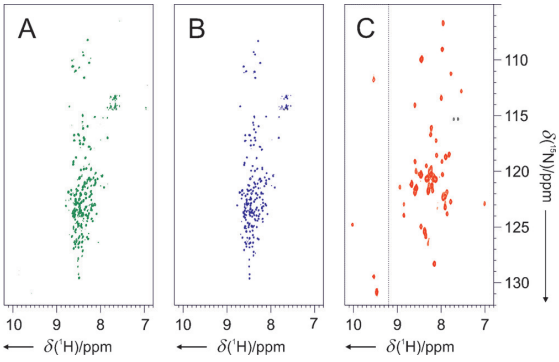


Figure 2. ^{1}H–^{15}N BEST-TROSY spectra recorded for E1A289 (panel A, green), E1A243 (panel B, blue), and the 2D ^{1}H-^{15}N SOFAST-HSQC spectrum recorded for E1ACR3 (panel C, red). In the latter panel, the left inset shows the same map drawn with a lower threshold for the sake of clarity. The complete chemical shift tables are deposited to the Biological Magnetic Resonance Data Bank (BMRB) with ID 25989, 25990, and 25991, respectively.

similar. Indeed, the number and distribution of the cross-peaks is almost equal, with minor differences in the intensity for some of the correlation peaks. In the spectrum of the longer form there is no evidence of the presence of signals due to the additional CR3 domain.

The tendency of E1A289 to form large aggregates upon increasing concentration poses limits in terms of protein concentrations that can be used for NMR experiments, restricting the choice among the available assignment strategies to those including the most-sensitive ^{1}H-detected NMR experiments. The E1A243 isoform instead doesn't show any tendency to aggregate, hence a higher protein concentration could be obtained, which allowed us to use ^{13}C direct-detected experiments for sequence specific assignment.[25,26] The analysis of experiments exploiting the higher chemical shift dispersion provided by the carbonyl dimension, enabled us to count 240 isolated cross peaks in the CON experiment, which corresponds to what is expected on the basis of the primary sequence, including 46 proline correlations. A series of 4D ^{13}C-detected experiments based on carbonyl detection were selected for the sequence-specific assignment[28–30] complemented by ^{1}H-detected experiments in the BEST-TROSY version,[23,31] providing the identification of 93% of the assignable backbone signals (BMRB 25989). In the case of E1A289, the acquired experiments were blind with respect to the CR3 domain, providing a less complete assignment, but still reporting about the conformation of 66% of the full length polypeptide (BMRB 25990). Information about the CR3 domain could be recovered by investigating the isolated CR3 domain. Its 2D ^{1}H–^{15}N spectrum (Figure 2C) shows a number of well distributed cross-peaks, consistent with the presence of secondary structural elements; the sequence-specific assignment of 97% of the CR3 domain could be obtained (BMRB 25991).

By comparing experimental chemical shifts of E1A243 and E1A289 with those predicted assuming a completely random coil conformation it is possible to identify partially populated secondary structural elements within E1A.[32,33] The resulting SSP scores, reported in Figure 3, show that in the N-terminal

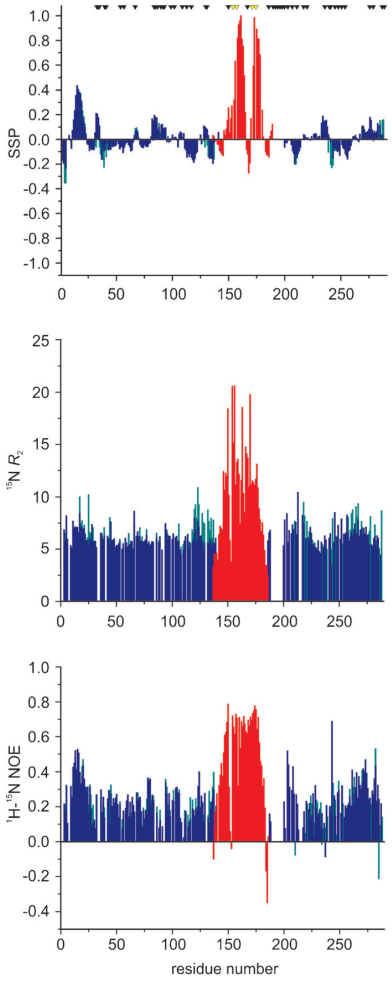


Figure 3. Secondary Structure Propensity (SSP) score and nuclear relaxation data for backbone amide ^{15}N obtained for E1A289 (green), E1A243 (blue), and E1ACR3 (red) from HAdV as a function of the residue number. From top to bottom, SSP, ^{15}N transverse relaxation rates (R_2, in Hz), and ^{1}H–^{15}N NOE values. The position of proline residues is indicated by black triangles, while the yellow triangles identify the Zn^{II}-binding cysteine residues (see text).

ChemPubSoc Europe

CHEMISTRY A European Journal Communication

part of the protein the amino acids stretch encompassing residues 12–22 has a significant α-helical secondary structural propensity, as suggested by previous studies on an N-terminal fragment of E1A.[21] The rest of the polypeptide does not show strong secondary structural propensities, with SSP values close to zero and small segments with SSP values within ±0.2. The results obtained for the detectable signals of the two variants are remarkably similar, showing that the two proteins in these regions are highly flexible and intrinsically disordered and share most of their structural and dynamic properties. As previously noted for other intrinsically disordered proteins,[34,35] prolines are very abundant (46 in total, that is, 16%) and appear to have an active role in modulating the structural and dynamic properties of the polypeptide chain: they are quite uniformly distributed throughout the primary sequence except for regions characterized by partially populated secondary structural elements. The protein is also very rich in negatively charged amino acids (Glu/Asp, 49 in total, that is, 17%), largely distributed in the disordered parts of the protein.

Heteronuclear relaxation rates, ^{15}N R_2 relaxation rates and ^{1}H–^{15}N heteronuclear NOEs in particular, provide useful information to characterize the dynamic properties of the protein. These data (Figure 3) indicate that the two E1A isoforms behave similarly for what concerns the common parts. Furthermore, the ^{1}H–^{15}N NOE values are significantly low, with an average value of 0.25, confirming that the proteins are largely disordered.

The detailed comparison of the data also shows some regions with meaningful differences in their heteronuclear relaxation properties, such as the regions flanking the CR3 and the CR4 region. The slightly higher ^{15}N R_2 relaxation rates observed for these parts of the protein in the E1A289 variant do indicate some interaction with the CR3 module present in E1A289 but the overall properties of the long disordered parts do not markedly change. A similar analysis was accomplished for E1ACR3. The SSP plot derived for the amino acids of this construct indicates the presence of two α-helices beginning with the CXXC motifs and including residues 153–165 and 170–180 (Figure 3). The five cysteine residues were assigned and the chemical shifts values for the C^{α} and C^{β} of Cys 154, 157, 171, and 174 are consistent with their involvement in Zn^{II}-binding. As this domain is not detectable in the E1A289 spectra we cannot confirm that these features are maintained in the full length protein. On the other hand, the results on the E1ACR3 clearly indicate the presence of a minimal fold composed by just two helices of less than 10 amino acids embedded in a relatively flexible polypeptide. The binding of Zn^{II} restricts the conformational space for the two helices, which should face each other to properly locate the ligand cysteines. Significantly higher values for the ^{1}H–^{15}N NOEs, confirm the presence of a more rigid domain in the short construct. However, the possibility to measure the proximity of these secondary structural elements is hindered by the low concentration of the sample, which should be maintained diluted due to the aggregation propensity of the CR3 domain. In fact, the analysis of 3D ^{15}N-edited and ^{13}C-edited NOESY experiments reveals only a few NOEs (essentially only intra-residue and sequential correlations) which are not sufficient to validate a structural model of this minimal fold. A statistical coil generator like *flexible-meccano*[36] can help visualizing the ensemble of possible conformations consistent with the information obtained so far. The pictures reported in Figure 4 provide a clear representation on how the different regions of E1A are able to expose the protein backbone for productive interactions with a number of partners, exploiting a number of SLiMs, and that the two partially structured regions (the first helix and the CR3 module) can adopt a multitude of different relative orientations.

The characterization presented here of the structural and dynamic properties of the two alternative spliced variants of E1A naturally exploited by HAdV nicely illustrates how intrinsic disorder is a key vehicle to encode in a relatively short polypeptide many functional modules that can be assembled to bring a tremendous functional complexity. The minimal Zn-binding module provided by the CR3 domain has been also character-

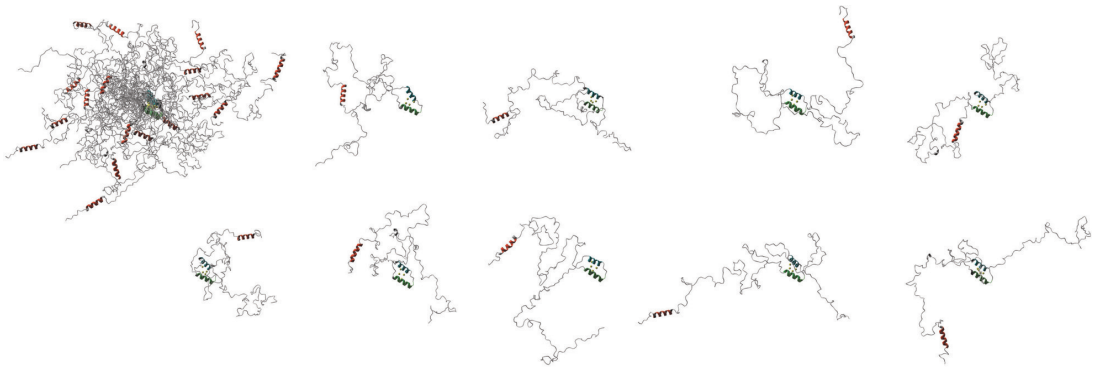


Figure 4. Conformers of E1A289 are shown to highlight the conformational heterogeneity of this protein. The protein is shown as a ribbon representation; the three regions characterized by α-helical secondary structural propensity are displayed in different colors (11–22 in red, 154–165 in green, and 170–180 in cyan). The sulfur atoms of the four cysteine residues forming the two CXXC motives (^{154}CXXC^{157} and ^{171}CXXC^{174}) are shown as yellow spheres. The panel on the left shows a subset of conformers superimposed using the CR3 domain coordinates (154–180). In addition, on the right some conformers are shown as examples of the many different relative orientations that the two more structured elements may adopt.

ChemPubSoc Europe

CHEMISTRY A European Journal Communication

ized. This unprecedented result, even if obtained on a small construct, represents a valid starting point to frame the many observations about its role in viral function.

Finally, many interactions have already been documented for the N-terminal fragment of E1A such as CBP/p300 and pRB, revealing peculiar ways for intermolecular interactions, and suggest a new concept of allostery modulated by intrinsic disorder.[37,38] The present results are going to provide a fertile ground to discover new ways in which proteins communicate in the cell and highlight key mechanisms through which hub proteins such as E1A are able to engage in an extremely complex network of intermolecular interactions and hijack cell regulation.

Experimental Section

All the details about cloning, expression and purification of the various isoforms of E1A, both ^{15}N and ^{13}C,^{15}N-labelled, studied in this work are reported in the Supporting Information. All the samples, with a concentration ranging between 75 and 500 μM, were in 10 mM HEPES buffer at pH 7.5, containing 50–150 mM KCl, and 1–10 mM DTT in H_2O with the addition of 10% D_2O for the lock signal. The ^{13}C-detected NMR experiments for sequence-specific assignment were acquired on a 16.4 T Bruker AVANCE 700 NMR spectrometer equipped with a cryogenically cooled triple resonance probehead optimized for ^{13}C direct detection. The ^{1}H-detected NMR experiments for sequence specific assignment were acquired on 21.1 T Bruker AVANCE 900 and 22.3 T Bruker AVANCE 950 NMR spectrometers, equipped with cryogenically cooled triple-resonance probeheads. ^{15}N relaxation experiments (R_1, R_2 and ^{1}H–^{15}N NOEs) were recorded on a 16.4 T Bruker AVANCE 700 spectrometer equipped with a cryogenically cooled triple resonance probehead. All experimental parameters used to acquire all the NMR experiments are reported in more detail in the Supporting Information.

Acknowledgements

This work has been supported by the European Commission Projects Marie Curie ITN IDPbyNMR (Contract No. 264257), BioNMR (Contract No. 261863) and INSTRUCT (Contract No. 211252). T.D.P. and M.O.N. are grateful to the program Science without Borders of the Brazilian Ministry of Science and Technology (CNPq) for financial support. Peter Pelka is gratefully acknowledged for providing the HAdV 2 plasmid while Alessandro del Grande, Alessandro Piai, Leonardo Gonnelli and Jochen Junker are gratefully acknowledged for many stimulating discussions.

Keywords: human adenovirus · intrinsically disordered proteins (IDP) · molecular hub · NMR spectroscopy · viral proteins

[1] S. M. Frisch, J. S. Mymryk, *Nat. Rev. Mol. Cell Biol.* **2002**, *3*, 441–452.
[2] P. Pelka, J. N. G. Ablack, G. J. Fonseca, A. F. Yousef, J. S. Mymryk, *J. Virol.* **2008**, *82*, 7252–7263.
[3] P. S. Moore, Y. Chang, *Nat.Rev.Cancer* **2010**, *10*, 878–889.
[4] J. J. Trentin, Y. Yabe, G. Taylor, *Science* **1962**, *137*, 835–841.
[5] S. M. Frisch, *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 9077–9081.
[6] S. Radko, R. Jung, O. Olanubi, P. Pelka, *Plos ONE* **2015**, *10*, e0140124.
[7] P. Hearing, T. Shenk, *Cell* **1983**, *33*, 695–703.
[8] J. N. G. Ablack, M. Cohen, G. Thillainadesan, G. J. Fonseca, P. Pelka, J. Torchia, J. S. Mymryk, *J. Virol.* **2012**, *86*, 8198–8209.
[9] P. Pelka, J. N. Ablack, M. Shuen, A. F. Yousef, M. Rasti, R. J. Grand, A. S. Turnell, J. S. Mymryk, *Virology* **2009**, *391*, 90–98.
[10] P. Pelka, M. S. Miller, M. Cecchini, A. F. Yousef, D. M. Bowdish, F. Dick, P. Whyte, J. S. Mymryk, *J. Virol.* **2011**, *85*, 8841–8851.
[11] P. Pelka, J. N. G. Ablack, J. Torchia, A. S. Turnell, R. J. Grand, J. S. Mymryk, *Nucleic Acids Res.* **2009**, *37*, 1095–1106.
[12] G. Chinnadurai, *Trends Microbiol.* **2011**, *19*, 174–183.
[13] N. Avvakumov, A. E. Kajon, R. C. Hoeben, J. S. Mymryk, *Virology* **2004**, *329*, 477–492.
[14] J. S. Culp, L. C. Webster, D. J. Friedman, C. L. Smith, W. J. Huang, F. Y. Wu, M. Rosenberg, R. P. Ricciardi, *Proc. Natl. Acad. Sci. USA* **1988**, *85*, 6450–6454.
[15] J. Habchi, P. Tompa, S. Longhi, V. N. Uversky, *Chem. Rev.* **2014**, *114*, 6561–6588.
[16] B. He, K. Wang, Y. Liu, B. Xue, V. Uversky, A. K. Dunker, *Cell Res.* **2009**, *19*, 929–949.
[17] E. Cilia, R. Pancsa, P. Tompa, T. Lenaerts, W. F. Vranken, *Nat. Commun.* **2013**, *4*, 2741.
[18] N. E. Davey, G. Trave, T. J. Gibson, *Trends Biochem. Sci.* **2011**, *36*, 159–169.
[19] K. Van Roey, B. Uyar, R. J. Weatheritt, H. Dinkel, M. Seiler, A. Budd, T. J. Gibson, N. E. Davey, *Chem. Rev.* **2014**, *114*, 6733–6778.
[20] P. E. Wright, H. J. Dyson, *Nat. Rev. Mol. Cell Biol.* **2015**, *16*, 18–29.
[21] J. C. Ferreon, M. A. Martinez-Yamout, H. J. Dyson, P. E. Wright, *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 13260–13265.
[22] R. Konrat, *J. Magn. Reson.* **2014**, *241*, 74–85.
[23] Z. Solyom, M. Schwarten, L. Geist, R. Konrat, D. Willbold, B. Brutscher, *J. Biomol. NMR* **2013**, *55*, 311–321.
[24] B. Brutscher, I. C. Felli, S. Gil-Caballero, T. Hošek, R. Kümmerle, A. Piai, R. Pierattelli, Z. Sólyom, *Adv.Exp.Med.Biol.* **2015**, *870*, 49.
[25] I. C. Felli, R. Pierattelli, *J. Magn. Reson.* **2014**, *241*, 115–125.
[26] S. Gil, T. Hošek, Z. Solyom, R. Kümmerle, B. Brutscher, R. Pierattelli, I. C. Felli, *Angew. Chem. Int. Ed.* **2013**, *52*, 11808–11812.
[27] I. C. Felli, R. Pierattelli (Eds) *Intrisically disordered proteins studied by NMR spectroscopy,* Springer, Heidelberg, **2015**, pp. 1–421.
[28] W. Bermel, I. Bertini, L. Gonnelli, I. C. Felli, W. Koźmiński, A. Piai, R. Pierattelli, J. Stanek, *J. Biomol. NMR* **2012**, *53*, 293–301.
[29] W. Bermel, I. C. Felli, L. Gonnelli, W. Koźmiński, A. Piai, R. Pierattelli, A. Zawadzka-Kazimierczuk, *J. Biomol. NMR* **2013**, *57*, 353–361.
[30] A. Piai, T. Hošek, L. Gonnelli, A. Zawadzka-Kazimierczuk, W. Koźmiński, B. Brutscher, W. Bermel, R. Pierattelli, I. C. Felli, *J. Biomol. NMR* **2014**, *60*, 209–218.
[31] E. Lescop, P. Schanda, B. Brutscher, *J. Magn. Reson.* **2007**, *187*, 163–169.
[32] K. Tamiola, F. A. Mulder, *Biochem. Soc. Trans.* **2012**, *40*, 1014–1020.
[33] M. Kjaergaard, F. M. Poulsen, *Prog. NMR Spectrosc.* **2012**, *60*, 42–51.
[34] P. Radivojac, L. M. Iakoucheva, C. J. Oldfield, Z. Obradovic, V. N. Uversky, A. K. Dunker, *Biophys. J.* **2007**, *92*, 1439–1456.
[35] A. Piai, E. O. Calçada, T. Tarenzi, A. del Grande, M. Varadi, P. Tompa, I. C. Felli, R. Pierattelli, *Biophys. J.* **2016**, *110*, 372–381.
[36] V. Ozenne, F. Bauer, L. Salmon, J. R. Huang, M. R. Jensen, S. Segard, P. Bernadó, C. Charavay, M. Blackledge, *Bioinformatics* **2012**, *28*, 1463–1470.
[37] P. Tompa, *Chem. Rev.* **2014**, *114*, 6715–6732.
[38] A. C. Ferreon, J. C. Ferreon, P. E. Wright, A. A. Deniz, *Nature* **2013**, *498*, 390–394.

Received: May 26, 2016
Published online on August 4, 2016

CHEMISTRY
A European Journal

Supporting Information

Structural and Dynamic Characterization of the Molecular Hub Early Region 1A (E1A) from Human Adenovirus

Tomáš Hošek[+], Eduardo O. Calçada[+], Marcela Oliveira Nogueira, Michele Salvi, Talita Duarte Pagani, Isabella C. Felli,* and Roberta Pierattelli*[a]

chem_201602510_sm_miscellaneous_information.pdf

Proteins expression and purification

HAdV E1A243

The gene codifying HAdV-2 E1A243 was optimized for the expression in *E.coli* and synthesized by GeneArt® (Invitrogen). AttL recognition sites were added to the gene at 3' and 5' to allow gateway recombination by LR clonase II enzyme mix into pETG-20A (from EMBL) destination vector. The resulting expression vector will generate E1A243 fused with thioredoxin A protein and a His-tag at the N terminus. The resulting expression vector, with the modified TEV cleavage site, yields the native 243 residues E1A243 protein with no additional amino acids.

The plasmid pETG-20A/HAdV-2 E1A243 was transformed into *E.coli* BL21(DE3) (Stratagene) and a single colony was selected to prepare a pre-inoculum in LB with ampicillin, grown overnight at 37°C, 180 rpm. The unlabelled, single labelled ^{15}N, and double labelled ^{13}C,^{15}N cultures were inoculated with 10 mL pre-inoculum into 1 L of chemically defined culture medium (48.5 mM Na_2HPO_4, 22.0 mM KH_2PO_4, 8.5 mM NaCl, 0.2 mM $CaCl_2$, 2.0 mM $MgSO_4$, 100 µM $ZnCl_2$, 1 mg/L each of biotin and thiamin, 7.5 mM $(NH_4)_2SO_4$ / $(^{15}NH_4)_2SO_4$ and 11.1 mM glucose / $^{13}C_6$-glucose) with ampicillin and grown at 37°C with constant agitation at 160 rpm. When OD_{600} reached 0.6-0.8, cells were induced for 5 h with IPTG (250 µM in case of unlabelled and single labelled ^{15}N cultures at 30°C; 100 µM in case of double labelled ^{13}C,^{15}N cultures at 37°C). The cells were harvested by centrifugation at 8000 *g* for 15 min and the cell pellet stored at -20°C.

All buffers used in E1A243 purification were deoxygenated and all the purification steps were taken inside an anaerobic chamber using refrigeration systems to keep the temperature lower than 10°C. Frozen cells were thawed and suspended in 25 mL of equilibration buffer A (50 mM HEPES pH 8.0, 500 mM NaCl, 5 mM imidazole, 2 mM DTT, 10 µM $ZnCl_2$ and Roche complete protease inhibitors). Cells were disrupted by sonication on ice (at 40% sonication amplitude using a Sonics Vibracell vcx 750 sonication system) with cycles of 3 s with 10 s delay pulses for 15 min. Lysate cells were centrifuged at 30000 rpm for 30 min at 4°C and the supernatant was loaded on a Ni^{2+} loaded HisTrap FF 5mL column (GE Healthcare Life Sciences) pre-equilibrated with buffer A. Column was washed using 10x column volumes of buffer A with 50 mM imidazole and the E1A243 protein was eluted using buffer A with 250 mM imidazole. The protein was concentrated to 2.5 mL and the buffer exchanged to 50 mM TRIS pH 8.0, 200 mM NaCl, 5 mM imidazole, 10 mM DTT using a PD10 column (GE Healthcare Life Sciences). 750 µL of TEV protease (0.1 mg/mL final conc.) were added for overnight digestion at room temperature inside the anaerobic chamber. After digestion, the E1A243 was further purified using Ni^{2+} loaded HisTrap FF 5ml column pre-equilibrated with buffer B (10 mM HEPES pH 8.0, 100 mM NaCl, 5 mM imidazole, 2 mM DTT). The digested E1A243 was eluted in the flow-through. The E1A243 was collected, warmed at 80°C for 20 minutes and centrifuged at 14000 *g* for 15 minutes. The supernatant was filtrated, concentrated at 5 mL and loaded on a HiLoad 16/100 Superdex 75 (GE Healthcare Life Sciences) size-exclusion chromatography column previously equilibrated with 10 mM HEPES pH 7.5, 50 mM KCl, 10 mM DTT using a flow-rate of 1.0 mL/min. The collected purified samples were checked by SDS-PAGE and mass-spectrometry to confirm the purity and molecular mass.

Protein concentration for NMR was estimated using the molar extinction coefficient (1490 $M^{-1}cm^{-1}$) calculated from ExPASy ProtParam tool [http://web.expasy.org/ protparam/]. The concentration was also confirmed using a Rudolph Research Analytical J357 automatic refractometer.

HAdV E1A289

The gene codifying HAdV-2 E1A289 construct inserted into the pET42a (EMD Millipore) expression vector in between the Nde1/Xho1 restriction sites, was generously provided by Peter Pelka. This construct contains two point mutations (S276P and E148G).

The plasmid pET42a-E1A289 was transformed into *E. coli* BL21(DE3)pLyS (Stratagene) and a single colony was selected to prepare a pre-inoculum in LB with kanamycin and chloramphenicol, grown overnight at 37°C, 180 rpm. The unlabelled, single labelled ^{15}N, and double labelled ^{13}C,^{15}N cultures were inoculated with 10 mL pre-inoculum into 1 L of chemically defined culture medium (48.5 mM Na_2HPO_4, 22.0 mM KH_2PO_4, 8.5 mM NaCl, 0.2 mM $CaCl_2$, 2.0 mM $MgSO_4$, 100 µM $ZnCl_2$, 1 mg/L each of biotin and thiamin, 7.5 mM $(NH_4)_2SO_4$ /$(^{15}NH_4)_2SO_4$ and 11.1 mM glucose / $^{13}C_6$-glucose) with kanamycin and chloramphenicol, grown at 37°C with constant agitation at 160 rpm. When OD_{600} reached 0.6-0.8, cells were induced with IPTG (500 µM in case of unlabelled and single labelled ^{15}N cultures; 100 µM in case of double labelled ^{13}C,^{15}N) for additional 8 h at 25°C. The cells were harvested by centrifugation at 8000 *g* for 15 min and the cells pellet stored at -20°C.

1

All buffers used for protein purification were deoxygenated and all the purification steps were taken inside an anaerobic chamber refrigerated to keep the temperature lower than 10°C. Frozen cells were thawed and suspended in 25 mL equilibration buffer A (50 mM HEPES pH 8.0, 500 mM NaCl, 5 mM imidazole, 2mM DTT, 10 µM $ZnCl_2$ and Roche complete protease inhibitors). Cells were disrupted by sonication on ice (at 40% sonication amplitude using a Sonics Vibracell vcx 750 sonication system) with cycles of 1 s with 10 s delay pulses for 30 min. Lysate cells were centrifuged at 30000 rpm for 30 min at 4°C and the supernatant was loaded on a 5 mL Ni^{2+} loaded HisTrap FF 5mL column (GE Healthcare Life Sciences) pre-equilibrated with buffer A. Column was washed using 10x column volumes of buffer A with 50 mM imidazole and the E1A289 protein was eluted using buffer A with 250 mM imidazole. The eluted sample was filtrated, concentrated at 5 mL and loaded on a HiLoad 16/100 Superdex 75 (GE Healthcare Life Sciences) size-exclusion chromatography column previously equilibrated with 10 mM HEPES, pH 7.5, 150 mM KCl, 10 mM DTT using a flow-rate of 1.0 mL/min. The collected purified samples were checked by SDS-PAGE and mass-spectrometry to confirm the purity and the molecular mass of the protein.

Protein concentration for NMR was estimated using the molar extinction coefficient (7450 $M^{-1}cm^{-1}$) calculated from ExPASy ProtParam tool [http://web.expasy.org/ protparam/]. The concentration was also confirmed by refractometer using a Rudolph Research Analytical J357 automatic refractometer.

HAdV E1ACR3

The gene codifying HAdV-2 E1ACR3 was synthesized by Eurofins Genomics in a pET21a (EMD Millipore) destination vector. The primary sequence contains residues 136-189 of the wild type protein and a methionine at the beginning of its sequence, comprising 54 residues of a native protein, without the addiction of a tag.

The plasmid was transformed into *E.coli* BL21(DE3) (Stratagene) and a single colony was selected to prepare a pre-inoculum in LB with ampicillin, grown overnight at 37°C, 180 rpm. 48 mL pre-inoculum were inoculated into 4 L of LB with ampicillin at 37°C, and as the cells reach OD_{600} 0.8, they were centrifuged gently at 4.500 rpm for 20 min at 4 °C, according to the Marley method[1]. The supernatant was discarded and the pellet was resuspended on a single labelled ^{15}N or double labelled ^{13}C,^{15}N media and the culture was inoculated into 1 L chemically defined culture medium (48.5 mM Na_2HPO_4, 22.0 mM KH_2PO_4, 8.5 mM NaCl, 0.2 mM $CaCl_2$, 2.0 mM $MgSO_4$, 100 µM $ZnCl_2$, 1 mg/L each of biotin and thiamin, 7.5 mM $(NH_4)_2SO_4$ /($^{15}NH_4)_2SO_4$ and 11.1 mM glucose / $^{13}C_6$-glucose) with ampicillin at 37°C with constant agitation at 180 rpm for 1 hour. After that, the culture was induced with 0.5mM IPTG, stored at 4 °C and at the end of the day the cells were shacked at 160 rpm, 17 °C for 16 h. The cells were harvested by centrifugation at 8000 *g* for 20 min and the cells pellet stored at -20°C.

All buffers used in protein purification were deoxygenated and all the purification steps were taken inside an anaerobic chamber using a refrigeration system to keep the temperature lower than 10 °C. Frozen cells were thawed and suspended 50 mL equilibration buffer A (20 mM TRIS, 10 mM NaCl pH 8.0, 1 mM DTT) and Roche complete mini protease inhibitor. Cells were disrupted by sonication on ice (at 80% sonication amplitude using a Sonics Vibracell vcx 750 sonication system) with cycles of 2 s with 10 s delay pulses for 25 min. Lysate cells were centrifuged at 30000 rpm for 30 min at 4 °C and the supernatant was filtered and loaded on a HiTrap Q FF 5 mL (Ge Healtcare Life Sciences), pre-equilibrated with buffer A. The recombinant E1ACR3 was eluted between 250 mM and 350 mM NaCl by a Buffer B (20 mM TRIS, 1 M NaCl, pH 8.0, 1 mM DTT) gradient from 0-50% in 15 minutes and 50-100% in 3 minutes. All the fractions containing the protein were concentrated until 5 mL and loaded on a HiLoad 16/60 Superdex 30 prep grade (GE Healthcare Life Sciences) size-exclusion chromatography column refrigerated at 4°C previously equilibrated with 10 mM HEPES pH 7.5, 50 mM KCl, 1 mM DTT using a flow-rate of 1.0 mL/min. The elution volume of the protein was about 80 mL. The fractions collected were conserved at 4°C for biophysical studies or frozen at -20 °C. The purified samples were checked by SDS-PAGE and mass-spectrometry to confirm the purity and the molecular mass of the protein. Zn(II) content was measured with an ICP Optical Emission Spectrometer Varian 720 and resulted to be 1:1.

Protein concentration was estimated by UV measurement using the molar extinction coefficient (5960 $M^{-1}cm^{-1}$) calculated from ExPASy ProtParam tool [http://web.expasy.org/protparam/] and confirmed by refractometer and with the Bradford assay.[2]

NMR Acquisition Parameters

The NMR experiments and acquisition parameters used for sequence specific assignment of H^N, N, C', C^α, $C^ß$ resonances and for further dynamics characterization were recorded at 278 K for HAdV E1A289 and E1A243 and at 288 K for E1ACR3.

The ^{13}C detected experiments for sequence specific assignment were acquired with a Bruker AVANCE 700 spectrometer equipped with a cryogenically cooled probe optimized for 13C direct detection (CON [3], 4D HCBCACON-[4], 4D HCBCANCO [4], 4D HCACONCACON [5] and 4D HNCONCACON [5]). The ^{13}C carrier was placed at 173 ppm, the 15N carrier at 124.8 ppm and the 1H carrier at 4.7 ppm. Band selective ^{13}C pulses were given at 173 and 39 ppm to excite or invert C' and C^{ali} spins respectively. The following band-selective pulses were used: 300 µs with newQ5 and time reversed newQ5 shapes for C' and C^{ali} excitation, 220 µs newQ3 shape for C'/C^{ali} inversion.

The ^{1}H detected experiments for sequence specific assignment were acquired with Bruker AVANCE 900 and Bruker AVANCE 950 spectrometers equipped with an inverse triple-resonance cryogenically cooled probehead (^{1}H-^{15}N BEST-TROSY [6;7], BEST-TROSY HNCO [8;9], BEST-TROSY HNCACO [8;9], BEST-TROSY HN(CO)CACB [8;9], BEST-TROSY HNCACB [8;9], BEST-TROSY HN(CA)NNH [8;10;11] and TROSY HN(COCA)NNH [12;13]). The ^{1}H carrier was placed at 4.7 ppm, the ^{15}N carrier at 118.5 ppm and the ^{13}C carrier at 173 ppm for C' and at 39 ppm for C^{ali}. Band selective ^{1}H pulses were given at 8.9 ppm to excite or invert H^N spins. The following band-selective pulses were used: 1800 µs with Pc9 shape for H^N excitation, 1270 µs EBURP and time reversed EBURP shapes for H^N excitation, 1250 µs REBURP shape for H^N inversion. Band selective ^{13}C pulses were given at 173 and 39 ppm to excite or invert C' and C^{ali} spins respectively. The following band-selective pulses were used: 274 µs with G4 and time reversed G4 shapes for C' and C^{ali} excitation, 190 µs Q3 shape for C'/C^{ali} inversion. The other experiments (^{15}N R_1, ^{15}N R_2 and ^{1}H-^{15}N NOEs) were acquired with a Bruker AVANCE 700 spectrometer equipped with an inverse triple-resonance cryogenically cooled probehead.

The conditions for E1A243 samples were 0.5 mM protein concentration in 10 mM HEPES buffer at pH 7.5, 50 mM KCl, and 10 mM DTT, for E1A289 were 0.2 mM in 10 mM HEPES buffer, pH 7.5, 150 mM KCl, 10 µM $ZnCl_2$, and 10 mM DTT and for E1ACR3 were 75 µM (unless otherwise specified) in 10 mM HEPES buffer, pH 7.5, 50 mM KCl and 1 mM DTT.

Experiments used for E1A289 sequence specific assignment	**Dimension of acquired data**			**Spectral width (ppm)**			**n[a]**	**d[b]**
	t_1	t_2	t_3	F_1	F_2	F_3		
^{13}C detected								
CON[c]	512* (^{15}N)	1024 (^{13}C)		38	30		48	2.2
HCBCACON[c]	128* (^{13}C)	64 (^{15}N)	1024 (^{13}C)	60	30	49	32	1
^{1}H detected								
^{1}H-^{15}N HSQC	256 (^{15}N)	1024 (^{1}H)		23	12		4	1
^{1}H-^{15}N BEST-TROSY	512 (^{15}N)	2048 (^{1}H)		30	12		8	0.2
BEST-TROSY HNCO	96 (^{13}C)	196 (^{15}N)	2048 (^{1}H)	8	24	15	8	0.2
BEST-TROSY HN(CA)CO	164 (^{13}C)	196 (^{15}N)	2048 (^{1}H)	8	24	15	16	0.2
BEST-TROSY HN(CO)CACB	164 (^{13}C)	196 (^{15}N)	2048 (^{1}H)	70	24	15	16	0.2
BEST-TROSY HNCACB	164 (^{13}C)	196 (^{15}N)	2048 (^{1}H)	70	24	15	16	0.2
BEST-TROSY HN(CA)NNH	156 (^{15}N)	156 (^{15}N)	2048 (^{1}H)	26	26	15	24	0.25
TROSY HN(COCA)NNH	96 (^{15}N)	150 (^{15}N)	2048 (^{1}H)	26	26	15	16	1
Dynamics								
^{15}N R_1[d]	350 (^{15}N)	2048 (^{1}H)		24	15		16	3.0
^{15}N R_2[d]	350 (^{15}N)	2048 (^{1}H)		24	15		16	3.0
steady-state heteronuclear ^{1}H -^{15}N-NOEs[d]	640 (^{15}N)	2048 (^{1}H)		24	15		48	6.0

[a] number of acquired scans.
[b] Relaxation delay in seconds.
[c] For experiments acquired in the IPAP mode, the dimension in which the two experiments are stored is indicated with an asterisk.
[d] In ^{15}N R_1, ^{15}N R_2 and heteronuclear ^{1}H-^{15}N-NOEs experiments the water signal was suppressed with the 'water flip-back' scheme. For the determination of R_1, 10 experiments were acquired changing the variable delay from 20 ms to 850 ms. For the determination of R_2, 9 experiments were acquired changing the variable delay from 18 ms to 250 ms.

Experiments used for E1A243 sequence specific assignment	Dimension of acquired data			Spectral width (ppm)			n[a]	d[b]
	t_1	t_2	t_3	F_1	F_2	F_3		
^{13}C detected								
CON-IPAP[c]	1024* (^{15}N)	1024 (^{13}C)		40	30		24	2
^{1}H detected								
^{1}H-^{15}N BEST-TROSY	512 (^{15}N)	2048 (^{1}H)		23	12		4	0.2
BEST-TROSY HNCO	164 (^{13}C)	196 (^{15}N)	2048 (^{1}H)	9.5	23	15	4	0.2
BEST-TROSY HN(CA)CO	164 (^{13}C)	196 (^{15}N)	2048 (^{1}H)	9.5	23	15	8	0.2
BEST-TROSY HN(CO)CACB	188 (^{13}C)	196 (^{15}N)	2048 (^{1}H)	58.1	23	15	8	0.2
BEST-TROSY HNCACB	188 (^{13}C)	196 (^{15}N)	2048 (^{1}H)	58.1	23	15	8	0.2
BEST-TROSY HN(CA)NNH	138 (^{15}N)	138 (^{15}N)	2048 (^{1}H)	23	23	15	24	0.28
TROSY HN(COCA)NNH	104 (^{15}N)	168 (^{15}N)	2048 (^{1}H)	23	23	15	8	1
Dynamics								
^{15}N R_1[d]	350 (^{15}N)	2048 (^{1}H)		24	15		16	3.0
^{15}N R_2[d]	350 (^{15}N)	2048 (^{1}H)		24	15		16	3.0
steady-state heteronuclear ^{1}H-^{15}N-NOEs[d]	640 (^{15}N)	2048 (^{1}H)		24	15		48	6.0

[a] number of acquired scans.
[b] Relaxation delay in seconds.
[c] For experiments acquired in the IPAP mode, the dimension in which the two experiments are stored is indicated with an asterisk.
[d] In ^{15}N R_1, ^{15}N R_2 and heteronuclear ^{1}H-^{15}N-NOEs experiments the water signal was suppressed with 'water flip-back' scheme. For the determination of R_1, 10 experiments were acquired changing the variable delay from 20 ms to 1 s. For the determination of R_2, 10 experiments were acquired changing the variable delay from 18 ms to 250 ms.

Experiment	Spectral width and maximal evolution times				n[a]	d[b]	cp[c]	hp[d]	t_{exp}[e]	ρ[f]
^{13}C detected										
4D HCBCACON	8800 Hz (^{13}C)	2600 Hz (^{15}N) 50ms	12500 Hz ($^{13}C^{ali}$) 7.6ms	5000 Hz (^{1}H) 10ms	8	0.9	512	850	35	0.07
4D HCBCANCO	8800 Hz (^{13}C)	2600 Hz (^{15}N) 32ms	12500 Hz ($^{13}C^{ali}$) 7.6ms	5000 Hz (^{1}H) 10ms	16	0.9	512	850	74	0.11
4D (HCA)CON(CA)CON	8800 Hz (^{13}C)	2600 Hz (^{15}N) 30ms	2200 Hz ($^{13}C'$) 25ms	2600 Hz (^{15}N) 25ms	16	0.9	512	930	82	0.34
4D (HN)CON(CA)CON	8800 Hz (^{13}C)	2600 Hz (^{15}N) 30ms	2200 Hz ($^{13}C'$) 25ms	2600 Hz (^{15}N) 25ms	16	0.5	512	910	54	0.33

[a] number of acquired scans.
[b] inter-scan delay in seconds.
[c] number of complex points.
[d] number of hypercomplex points.
[e] duration of the experiment.
[f] relative data points density (%)

Experiments used for E1ACR3 sequence specific assignment	Dimension of acquired data			Spectral width (ppm)			n[a]	d[b]
	t_1	t_2	t_3	F_1	F_2	F_3		
^{1}H detected								
^{1}H-^{15}N HSQC	128 (^{15}N)	2048 (^{1}H)		28	16		8	1.5
^{1}H-^{15}N BEST-TROSY	256 (^{15}N)	3072 (^{1}H)		28	12		8	0.3
BEST-TROSY HNCO	140 (^{13}C)	96 (^{15}N)	3072 (^{1}H)	8	28	12	8	0.3
CBCACONH[e]	156 (^{13}C)	96 (^{15}N)	2048 (^{1}H)	55	28	12	16	1

BEST-TROSY HN(CO)CACB	180 (^{13}C)	96 (^{15}N)	3072 (^{1}H)	70	28	12	16	0.3
BEST-TROSY HNCACB	180 (^{13}C)	96 (^{15}N)	3072 (^{1}H)	70	28	12	16	0.3
BEST-TROSY HNCA[e]	140 (^{15}N)	96 (^{15}N)	3072 (^{1}H)	28	28	12	40	0.3
BEST-TROSY HN(CO)CA[e]	140 (^{15}N)	96 (^{15}N)	3072 (^{1}H)	28	28	12	32	0.3
Dynamics								
^{15}N R_1[d]	128 (^{15}N)	1800 (^{1}H)		28	16		32	3.0
^{15}N R_2[d]	128 (^{15}N)	1800 (^{1}H)		28	16		40	3.0
steady-state heteronuclear ^{1}H-^{15}N-NOEs[d]	256 (^{15}N)	1800 (^{1}H)		28	16		152	6.0

[a] number of acquired scans.
[b] Relaxation delay in seconds.
[c] For experiments acquired in the IPAP mode, the dimension in which the two experiments are stored is indicated with an asterisk.
[d] In ^{15}N R_1, ^{15}N R_2 and heteronuclear ^{1}H-^{15}N-NOEs experiments the water signal was suppressed with 'water flip-back' scheme. For the determination of R_1, 10 experiments were acquired changing the variable delay from 20 ms to 1000 ms. For the determination of R_2, 10 experiments were acquired changing the variable delay from 18 ms to 250 ms.
[e] Sample concentration was 140 μM

NMR data processing and analysis

Conventionally sampled NMR data sets were processed using either *Bruker TopSpin 1.3* or *TopSpin 2.0* software. Instead, when non uniform sampling (NUS) was employed, the NMR data were converted with *nmrPipe* [14] and then processed using either the Multidimensional Fourier Transform (MFT) algorithm (for 3D data sets) or the Sparse MFT (SMFT) algorithm (for 4/5D data sets), respectively, as implemented in *ToASTD* [15] and in *reduced* [16;17] programs (available at http://nmr.cent3.uw.edu.pl). CcpNMR [18], *CARA* [19] and *Sparky* [20] were used to analyze and annotate the spectra.

The ^{15}N relaxation rates (R_1 and R_2) were determined by fitting the cross-peak intensity measured as a function of variable delay, to single-exponential decay. ^{1}H-^{15}N NOE values were obtained as a ratio between peak intensity in spectra recorded with and without ^{1}H saturation.

The secondary structure propensity (SSP) from heteronuclear chemical shifts was determined by using the neighbor corrected structural propensity calculator (ncSPC) tool [21] available online at http://nmr.chem.rug.nl. The Tamiola, Acar and Mulder random coil chemical shift library [22] was chosen for the analysis.

To obtain the ensemble description of E1A289, 5,000 conformers were generated using the program *flexible-meccano* [23], available at http://www.ibs.fr. The SSP score calculated for the first helix of E1A289 was used as input in the calculation while the regions encompassing residues 153-164 and 170-180 were imposed to have helical conformations assigning to each residue an SSP value equal to 1; the ^{154}CXXC^{157} and ^{171}CXXC^{174} fragments were further constrained to be at contact distance (less than 10 Å), imposing a further restrain in the calculation.

Supplementary Figures

Figure S1

Nuclear relaxation data for backbone amide ^{15}N for (from top to bottom) HAdV E1A289 (green), E1A243 (blue) and E1ACR3 (red) as a function of the residue number. From left to right, ^{15}N longitudinal relaxation rates (R_1, Hz), ^{15}N transverse relaxation rates (R_2, Hz), and ^{1}H-^{15}N NOE values. Experiments were recorded on a 16.4 T NMR spectrometer as reported above.

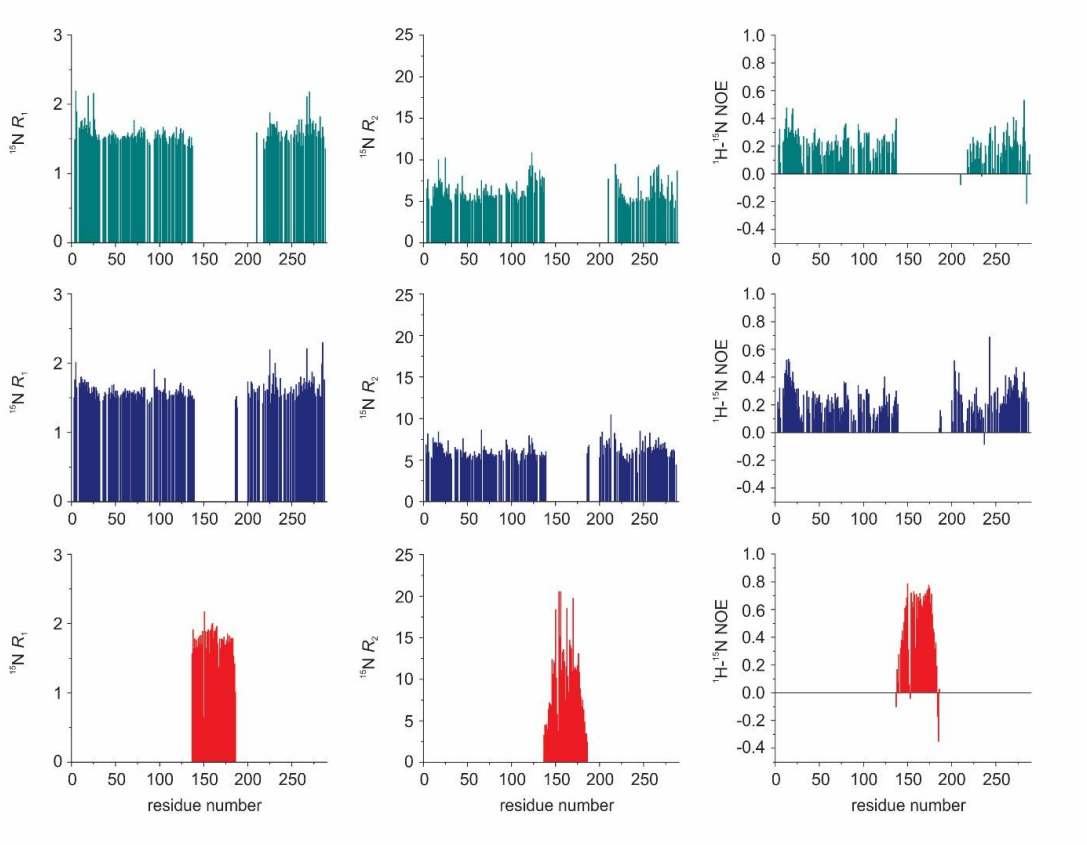
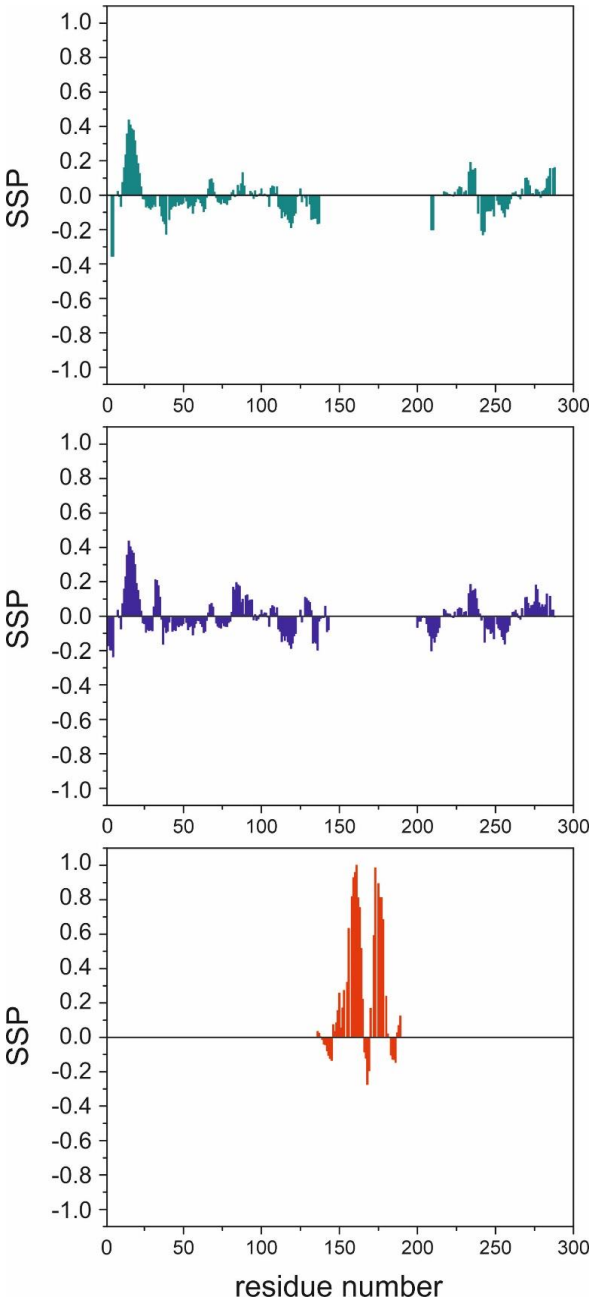


Figure S2

Secondary structure propensity (SSP) score[21] obtained by combining chemical shifts for the all assigned resonances for (from top to bottom) HAdV E1A289 (green), E1A243 (blue) and E1ACR3 (red).



References

[1.] J. Marley, M. Lu, C. Bracken, *Journal of Biomolecular NMR* **2001,** *20* 71-75.

[2.] M. Bradford, *Anal.Biochem.* **1976,** *72* 248-254.

[3.] W. Bermel, I. Bertini, L. Duma, L. Emsley, I. C. Felli, R. Pierattelli, P. R. Vasos, *Angew.Chem.Int.Ed.* **2005,** *44* 3089-3092.

[4.] W. Bermel, I. Bertini, L. Gonnelli, I. C. Felli, W. Kozminski, A. Piai, R. Pierattelli, J. Stanek, *J.Biomol.NMR* **2012,** *53* 293-301.

[5.] W. Bermel, I. C. Felli, L. Gonnelli, W. Kozminski, A. Piai, R. Pierattelli, A. Zawadzka-Kazimierczuk, *J.Biomol.NMR* **2013,** *57* 353-361.

[6.] A. Favier, B. Brutscher, *J.Biomol.NMR* **2011,** *49* 9-15.

[7.] E. Lescop, T. Kern, B. Brutscher, *J.Magn.Reson.* **2010,** *203* 190-198.

[8.] E. Lescop, P. Schanda, B. Brutscher, *J.Magn.Reson.* **2007,** *187* 163-169.

[9.] M. Salzmann, K. Pervushin, G. Wider, H. Senn, K. Wüthrich, *Proc.Natl.Acad.Sci.USA* **1998,** *95* 13585-13590.

[10.] P. Schanda, H. Van Melckebeke, B. Brutscher, *J.Am.Chem.Soc.* **2006,** *128* 9042-9043.

[11.] R. Weisemann, H. Rüterjans, W. Bermel, *J.Biomol.NMR* **1993,** *3* 113-120.

[12.] C. Bracken, A. G. Palmer, III, J. Cavanagh, *J.Biomol.NMR* **1997,** *9* 94-100.

[13.] S. C. Panchal, N. S. Bhavesh, R. V. Hosur, *J.Biomol.NMR* **2001,** *20* 135-147.

[14.] F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer, A. Bax, *J.Biomol.NMR* **1995,** *6* 277-293.

[15.] K. Kazimierczuk, A. Zawadzka, W. Kozminski, I. Zhukov, *J.Biomol.NMR* **2006,** *36* 157-168.

[16.] K. Kazimierczuk, A. Zawadzka, W. Kozminski, *J.Magn.Reson.* **2009,** *197* 219-228.

[17.] K. Kazimierczuk, A. Zawadzka-Kazimierczuk, W. Kozminski, *J.Magn.Reson.* **2010,** *205* 286-292.

[18.] W. F. Vranken, W. Boucher, T. J. Stevens, R. H. Fogh, A. Pajon, Llinas M, E. L. Ulrich, J. L. Markley, J. Ionides, E. D. Laue, *Proteins: Struct., Funct., Bioinf.* **2005,** *59* 687-696.

[19.] Keller, R. L. J. The Computer Aided Resonance Assignment Tutorial. Cantina Verlag, **2004**.

[20.] Goddard, T. D. and Kneller, D. G. SPARKY 3, University of California, San Francisco, **2000**.

[21.] K. Tamiola, F. A. Mulder, *Biochem.Soc.Trans.* **2012,** *40* 1014-1020.

[22.] K. Tamiola, B. Acar, F. A. A. Mulder, *J.Am.Chem.Soc.* **2010,** *132* 18000-18003.

[23.] V. Ozenne, F. Bauer, L. Salmon, J. R. Huang, M. R. Jensen, Segard S., P. Bernadó, Charavay C., M. Blackledge, *Bioinformatics* **2012,** *28* 1463-1470.

Section 3.2

Toward the real-time monitoring of HPV-16 E7 phosphorylation events

Marcela O. Nogueira[a], Tomáš Hošek[a], Eduardo O. Calcada[a], Francesca Castiglia[a], Paola Massimi[b], Lawrence Banks[b], Isabella C. Felli[a], Roberta Pierattelli[a]

[a] *Magnetic Resonance Center (CERM) and Department of Chemistry "Ugo Schiff", University of Florence, Italy*

[b] *International Centre for Genetic Engineering and Biotechnology (ICGEB),, Trieste, Italy*

Submitted to Virology

3.2. TOWARD THE REAL-TIME MONITORING OF HPV-16 E7 PHOSPHORYLATION EVENTS

ABSTRACT

HPV-16 E7 is one of the key proteins that, by interfering with the host metabolism through many protein-protein interactions, hijack cell regulation and contributes to malignancy. Here we report the high-resolution investigation of the CR3 region of HPV-16 E7, both as an isolated domain and in the full-length protein. This opens the way to the atomic level study of the many interactions in which HPV-16 E7 is involved. Along these lines we show here the effect of one of the key post-translational modifications of HPV-16 E7, the phosphorylation by casein kinase II.

INTRODUCTION

Human papillomaviruses (HPVs) constitute a family of more than 100 viruses, which infect the mucosa and the squamous epithelia [1]. HPVs are classified as Low risk or high Risk depending on their propensity to cause either benign or dysplastic lesions which can ultimately lead to cancer [1–3]. Within the HR group, HPV-16 and HPV-18 are the types most often found in cervical cancer, which is the second most common cancer in women worldwide [2,4].

HPVs have a small circular double stranded DNA (about 8 kb) encoding only a few proteins. Among them, two proteins, E6 and E7, play an important role in cancer development in the oncogenic variants [5]. Both proteins work in cooperation to hijack cell regulation by interacting with many proteins of the host, such as the tumor suppressors p53, in case of E6, and pRb in case of E7 [6–8]. However, the presence of E7 is sufficient to immortalize epithelial cells even in absence of E6 [9] and in transgenic mouse models E7 plays a major role in cervical cancer development [10]. This shows the importance of E7 in the development of HPV-related malignancy caused by HPVs.

HPV-16 E7 is a heterogeneous protein in terms of its structural and dynamic properties. It is 98 amino acids long, with three conserved regions (CR) [11]. CR1 and CR2 are located in the N-terminal half, which is intrinsically disordered and is characterized by high flexibility [12,13]. CR3 is in the C-terminal part of the protein, which appears to be more structured and which contains two CXXC zinc-binding motifs separated by 29 amino acids (Figure 1A). This domain is also responsible for

the formation of a homodimer, the form in which E7 is was found in other variants *in vitro* [13–15].

E7 acts as a hub protein, interacting with several partners that are involved in many different processes, such as cell cycle control and apoptosis [16–20]. One of the most important and well-described interactions for HPV-16 E7 is with the retinoblastoma tumor suppressor (pRb), which binds through the LXCXE motif present in CR2 (Figure 1A) [8,21]. However, optimal interaction with pRb also requires residues in the CR3 domain [22–24]. Indeed, CR3 is involved in many other interactions that contribute to HPV-mediated oncogenesis such as, for example, interactions with the cyclin-dependent kinase inhibitors p21 and p27, and transcription factors like the TATA box-binding protein (TBP), p300/CBP and E2F. Some of these interactions are strongly modulated by post-translational modifications (PTMs) of E7. In particular, the phosphorylation of Ser31 and Ser32 by casein kinase II (CKII) has been shown to play an important role in modulating many of E7s activities. For example CKII phosphorylation of E7 plays an important role in the ability of E7 to transform cells and for its ability to interact with TBP [20,25]. Recent studies also suggest that the acidic domain that comprises the CKII recognition site, which lies downstream of the LXCXE motif, also contributes towards the recognition of pRb [26,27]. Furthermore an intact CKII phosphorylation site was also found to be important for E7 to be able to target the pRb and the related p130 pocket proteins for degradation [28,29]. However, what effects phosphorylation of E7 by CKII has upon the structural and dynamic properties of E7 is still unknown. Whilst recent studies have used nuclear magnetic resonance (NMR) spectroscopy to characterize the N-terminal domain of HPV-16 E7 [12] and structures of the C-terminal region have been reported for HPV-1a [14] and HPV-45 E7 [13], there is currently no information on how phosphorylation of E7 by CKII might modulate this structure.

Here we describe the atomic resolution characterization of E7CR3 from HPV-16. Moreover we investigated the effect of phosphorylation of E7 through time-resolved NMR experiments. This provides the basis for detailed studies of the properties of the phosphorylated forms of E7 and how this modulates its interactions with many cellular proteins.

RESULTS AND DISCUSSION

For any detailed analysis of E7 at the molecular level, including the effect of post-translational modifications (PTMs) and protein-protein interactions, it is necessary to have the sequence-specific assignment of the polypeptide. The NMR analysis of the full-length protein had provided only the assignment of the N-terminal portion, comprising the CR1 and the CR2 regions [12], since the weak intensity and the broadening of the signals from the CR3 region in the 2D and 3D NMR spectra of E7 did not allow detection of the correlations necessary to achieve sequence-specific assignment. Therefore, the isolated E7 CR3 domain was expressed, purified and analysed by NMR spectroscopy.

The overlay of the ^{1}H-^{15}N correlation spectra acquired from the full length protein (E7) with that from the CR3 domain (E7CR3) clearly shows that the structural properties of the CR3 domain are quite similar in both polypeptides, since there are only minor changes in the chemical shift between the two spectra (Figure 1B). This also means that the assignment obtained for the E7CR3 can be easily transferred to the full length E7.

On the shorter construct a series of 3D measurements could be acquired with sufficient sensitivity and resolution to achieve the sequence-specific assignment of the majority of the E7 CR3 signals. A few cross peaks in the 2D ^{1}H-^{15}N correlation spectra, most likely deriving from the 64-72 fragment, could not be assigned because of insufficient information in the 3D spectra. Chemical shifts of $^{1}H^{N}$, ^{15}N, $^{13}C'$, $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ of the assigned residues are reported in Table 1 and deposited in the BioMagResBank (accession number 26069 at http://www.bmrb.wisc.edu/). The assignment of the signals of E7 CR3 was then transferred to those of CR3 in the full-length protein.

The chemical shifts obtained from the sequence-specific assignment were used for the secondary structure propensity (SSP) analysis [30]. In Figure 2 the SSP score for the initial disordered N-terminal half of E7 is also shown. The results are in agreement with the structural investigations on the other HPV serotypes [13,14], and confirm the high propensity of the initial N-terminal part of CR3 (residue 45-58) to

adopt an extended conformation, and also confirm the presence of two α-helixes in the regions from residue 69 to 82 and 92 to 96.

The assignment of $^{1}H^{N}$ and ^{15}N resonances of about 90% of the amino acids of HPV-16 E7 allows us to monitor at atomic resolution the occurrence of PTMs that are linked to its high oncogenic potential.

As mentioned above, E7 is a heterogeneous protein in terms of structural and dynamic properties. The folded domain (CR3) has different spectral properties from the flexible domain. The former yields quite weak but well dispersed signals, while the latter yields very intense signals clustered in a very narrow region. Thus, in order to follow both sets of signals, we acquired two different variants of 2D ^{1}H-^{15}N correlation experiments, one optimized for sensitivity (^{1}H-^{15}N SOFAST-HMQC) to monitor the signals from CR3 and one optimized for resolution (^{1}H-^{15}N HSQC) to follow the signals deriving from the highly disordered and flexible N-terminal part (CR1 and CR2). This strategy allowed us to monitor the majority of signals of E7 despite their extremely different properties.

Several interactions of E7 are modulated by the phosphorylation of E7 by Casein Kinase II (CKII) [31,32,24,33], which phosphorylates E7 in the CR2 region, recognising the specific motif, **SS**EEEDE [22,29]. Therefore, we decided to investigate the effect of this post-translational modification on the structure of the full-length E7.

Before monitoring the in vitro CKII phosphorylation of E7 by NMR, the conditions were optimized using autoradiography essays, in particular to monitor the effect of the buffer and the reducing agent present in the buffer. Afterwards, a series of time-resolved NMR experiments were performed to monitor the complete phosphorylation of the protein.

During the time-course of the reaction the appearance of NMR signals and the disappearance of others was observed. The signals have the characteristic ^{1}H signals from the backbone amides of phosphorylated serine residues expected at higher chemical shifts (Figure 3A) [34]. The evolution of the signal intensity of the two ^{1}H-^{15}N backbone correlations of un-phosphorylated and phosphorylated serine residues is reported in Figure 3B.

This series of time-resolved experiments enables us also to analyze in detail the chemical shift and intensity changes of the signals throughout the reaction to assess whether only Ser31 and Ser32 are affected, or whether other residues are also perturbed, either located close or distant from the phospho-acceptor sites.

The chemical shifts of backbone amide resonances for most of E7 residues are unchanged upon Ser31 and Ser32 phosphorylation (Figure 3A). Only amino acids closer to the phosphorylated serine residues in the primary sequence (in particular residues 26-29) are significantly affected by the phosphorylation reaction, as would be expected (Figure 3A). Interestingly this potentially affects the pRb binding LXCXE motif, which spans amino acid residues 22-26 [35]. Overall, the fingerprint of the E7CR3 domain (well dispersed signals in the 2D ^{1}H-^{15}N correlation map) is maintained throughout the phosphorylation reaction, suggesting that this domain remains stable, in terms of structure and conformation, throughout the reaction.

At the end of the reaction, the phosphorylated E7 was isolated from the medium to evaluate the stability of the phosphorylated form of the protein. Once the buffer was exchanged to eliminate the reagents, E7 was shown to be stable for more than one week. The overall maintenance of the signal intensities also suggests that E7 does not form big oligomers after the phosphorylation reaction as has been suggested in previous studies [36]. These findings open the possibility for further atomic resolution investigation of how phosphorylation of the protein influences its properties as well as its interactions with protein partners.

CONCLUSIONS

The sequence-specific assignment of E7CR3 represents an important additional step in the characterization of HPV-16 E7. The structural properties of the CR3 domain are quite similar in the full-length E7 and the isolated CR3 region and the assignment obtained from the isolated domain could be transferred to the full-length protein. This information, essential for better understanding the peculiarities of this protein, could thus be used to investigate the impact of one of the most important PTMs on E7, the CKII-dependent phosphorylation occurring in two adjacent serine residues. In the present work we have optimized the experimental conditions to study the

phosphorylation of E7 *in vitro* in a time-resolved manner through NMR spectroscopy. With these investigations we were able to monitor all signals from HPV-16 E7 during the phosphorylation of Ser31 and Ser32 by CKII, and we can finally conclude that the phosphorylation of both serines does not affect E7's overall conformation, which remains stable several days after the reaction. Previous studies have however shown an important requirement for the acidic residues that make up the CKII consensus motif for optimal pRb recognition [26,27], and presumably the negative charges added by phosphorylation may also play an additional role in this [28,29]. However our studies here also highlight the chemical shifts in backbone residues as far away from the CKII phospho-acceptor site as E26, which is an integral part of the pRb recognition motif [37]. This suggests that phosphorylation of E7 by CKII may also modulate the LXCXE/pocket protein interaction. Taken together these results suggest that alterations in E7's capacity to interact with its target proteins as a result of CKII phosphorylation are most likely all due to local changes in charge at Ser31/Ser32 and minor changes to the local residues 26-29, and not due to major structural alterations in the rest of the E7 protein.

MATERIALS AND METHODS

Full-length E7 expression and purification

For the expression of E7 wild type protein from HPV-16 (amino acids 1-98) the recombinant vector pET20-E7-opt [12] containing the gene for expression of E7 protein (1-98-His_6), was mutated by the insertion of a stop codon (TAG) between the codon for Pro 98 and the codons for the His_6 tag, generating the pET20-E7-WT. Thus, that plasmid was used for transformation of *E. coli* BL21 pLys (Stratagene®) cells.

After expression tests, one colony was selected for the growing in LB medium with ampicillin (0.1 mg/mL) and chloramphenicol (0.34 mg/mL). Finally, the cells were cultured in 1 L of M9 medium at 37 °C containing $(^{15}NH_4)_2SO_4$, 0.1 mg/mL ampicillin, 0.34 mg/mL chloramphenicol and 100 μM $ZnCl_2$. When the culture reached OD_{600} 0.7 protein expression was induced by the addition of IPTG (isopropyl β-D-1- thiogalactopyranoside) during 5h at 37 ºC. After that the cells were harvested at 4.500 rpm for 20 min at 4°C. The supernatant was discarded and the cells were re-

suspended in 25 mL of degassed buffer A (20 mM TRIS buffer at pH 8.5, 10 mM DTT supplemented with Roche complete protease inhibitors). After this the bacterial cells were disrupted by sonication (20 kHz sonication frequency and 40% of amplitude) with cycles of 3s and 10s of delay in anaerobic conditions. Lysed cells were then centrifuged at 165000 *g*. The supernatant containing E7 protein was loaded on a HiTrap QFF 10 mL ionic exchange column pre-equilibrated with buffer A. The purification step was followed with buffer A in a gradient with buffer B (buffer A + 0.5 M NaCl). E7 protein was eluted with 0.5 M NaCl (100% buffer B). Fractions containing the protein were mixed, concentrated to 4 mL and injected in a Superdex 75 16/100 size exclusion chromatography column pre-equilibrated with the final degassed buffer (10 mM HEPES pH 7.5, 50 mM KCl, 10 mM DTT 10 μM $ZnCl_2$). After analysis by SDS-PAGE 17%, fractions containing the pure protein were mixed and concentrated to 500 μL for NMR analysis. 50 μL of D_2O was added to each NMR sample for the lock signal. Zn^{2+} content was measured by Inductively Coupled Plasma-Optical Emission Spectrometer (ICP-OES) Varian 720 and resulted to be 2:1 (E7:Zn^{2+}).

Expression and purification of E7CR3 domain

E. coli Bl21 (DE3) (Stratagene®) were transformed with the plasmid pET21a-E7-CR3_WT (Invitrofins®), encoding the gene to express the E7CR3 construct comprising amino acids 45-98. After expression tests, one colony was selected to grow in LB medium with 0.1 mg/mL of ampicillin and 0.34 mg/mL of chloramphenicol. Finally, cell cultures were grown in 1 L of M9 medium at 37 °C enriched with $(^{15}NH_4)_2SO_4$, $^{13}C_6$-glucose and 100 μM $ZnCl_2$. When the culture reached OD_{600} of 0.6-0.7, cells were induced with 250 μM IPTG and allowed to grow for 5 additional hours at 37 ºC. Thus, cells were harvested by centrifugation at 8000 *g* for 15 minutes at 4 ºC.

HPV-16 E7 CR3 was expressed in inclusion bodies, which were suspended with degassed 50 mM HEPES buffer, pH 8.5, 50 mM KCl with 2 M urea in anaerobic conditions. The solubilized protein was concentrated to 4 mL and injected in a Superdex 75 16/100 size exclusion chromatography column pre-equilibrated with the

final degassed buffer (50 mM HEPES pH 8.5, 50 mM KCl, 5 mM DTT, 10 μM $ZnCl_2$). Fractions containing the pure protein were mixed and concentrated to 500 μL (250 μM concentration) for NMR experiments.

E7 phosphorylation by CKII

The purified full-length HPV-16 E7 protein was *in vitro* phosphorylated in suspension buffer (10 mM HEPES pH 7.5, 10 mM DTT, 10 μM $ZnCl_2$), with 20 mM $MgCl_2$, 25μM [^{32}P] γ-ATP (in the control), 0.3 μM Aprotinin, 1 μM Pepstatin, 50 mM NaF, 4 mM Na_3VO_4, 200 Units of recombinant human CKII kinase enzyme (New England Biolabs) at 30°C for 20 minutes. The reaction was monitored by SDS-gel and autoradiography [32].

For the NMR experiments, the protein was treated with different concentrations of Mg-ATP from Sigma-Aldrich® and also different concentration of recombinant human CKII from New England Biolabs® as reported in the main text. The reagents were added directly in the NMR tube containing the protein in the suspension buffer.

NMR experiments

Sequence specific assignment of E7CR3 from HPV-16

The ^{1}H detected NMR experiments for sequence specific assignment were acquired at 298 K on the 22.3 T Bruker AVANCE 950 spectrometer, operating at 950.2 MHz for ^{1}H, equipped with cryogenically cooled triple resonance probehead. A dataset of ^{1}H detected NMR experiments including ^{1}H-^{15}N BEST-TROSY [38,39], ^{1}H-^{13}C HSQC [40–42], BEST-TROSY HNCO [43,44], BEST-TROSY HNCA [43,44], CBCACONH [45], HCCH-TOCSY [46] were acquired for sequence specific assignment of H^N, N, C', C^α, $C^ß$ resonances.

The ^{1}H carrier was placed at 4.7 ppm, the ^{15}N carrier at 121.3 ppm and the ^{13}C carrier at 173 ppm for C' and at 39 ppm for C^{ali}. Band selective ^{1}H pulses were given at 8.9 ppm to excite or invert H^N spins. The following band-selective pulses were used: 1800 μs with Pc9 shape for H^N excitation, 1270 μs EBURP and time reversed

EBURP shapes for H^N excitation, 1250 μs REBURP shape for H^N inversion [47]. Band selective ^{13}C pulses were given at 173 and 39 ppm to excite or invert C' and C^{ali} spins respectively. The following band-selective pulses were used: 274 μs with G4 and time reversed G4 shapes for C' and C^{ali} excitation, 190 μs Q3 shape for C'/C^{ali} inversion.

Data were processed with Bruker TopSpin 2.0 and analyzed with ccpNMR [48] and Computer Aided Resonance assignment (CARA) [49] programs. The secondary structure propensity was determined from chemical shifts data by using the ncSPC web-portal [http://nmr.chem.rug.nl/ncSPC/] [30].

Time resolved experiments to follow E7 phosphorylation

In vitro E7 phosphorylation was followed with series of NMR experiments (1D ^1H NMR, ^1H-^{15}N SOFAST-HMQC [50,51] and ^1H-^{15}N HSQC [40–42,52]) acquired over the time of the reaction. All NMR experiments were acquired at 303 K on a 16.7 T Bruker AVANCE 700 MHz NMR spectrometer equipped with a cryogenically cooled triple resonance probehead. ^1H-^{15}N SOFAST-HMQC experiments were recorded with 16 transients and 1204 (^1H) x 128 (^{15}N) complex points. ^1H-^{15}N HSQC were recorded with 4 transients and 1536 (^1H) x 256 (^{15}N) complex points. Each experiment required about 20 minutes of acquisition time. NMR data were processed with TopSpin 2.0 and analyzed with the program ccpNMR [48].

ACKNOWLEDGMENTS

This work has been supported by the European Commission Projects Marie Curie ITN IDPbyNMR (Contract No. 264257), BioNMR (Contract No. 261863) and INSTRUCT (Contract No. 211252). LB and PM gratefully acknowledge research support provided by the Associazione Italiana per la Ricerca sul Cancro (Grant number 14025). MON is grateful to the program Science without Borders of the Brazilian Ministry of Science and Technology (CNPq) for financial support. Jochen Junker and Leonardo Gonnelli are gratefully acknowledged for many stimulating discussions. We are also grateful to Miranda Thomas for valuable comments on the manuscript.

FIGURES

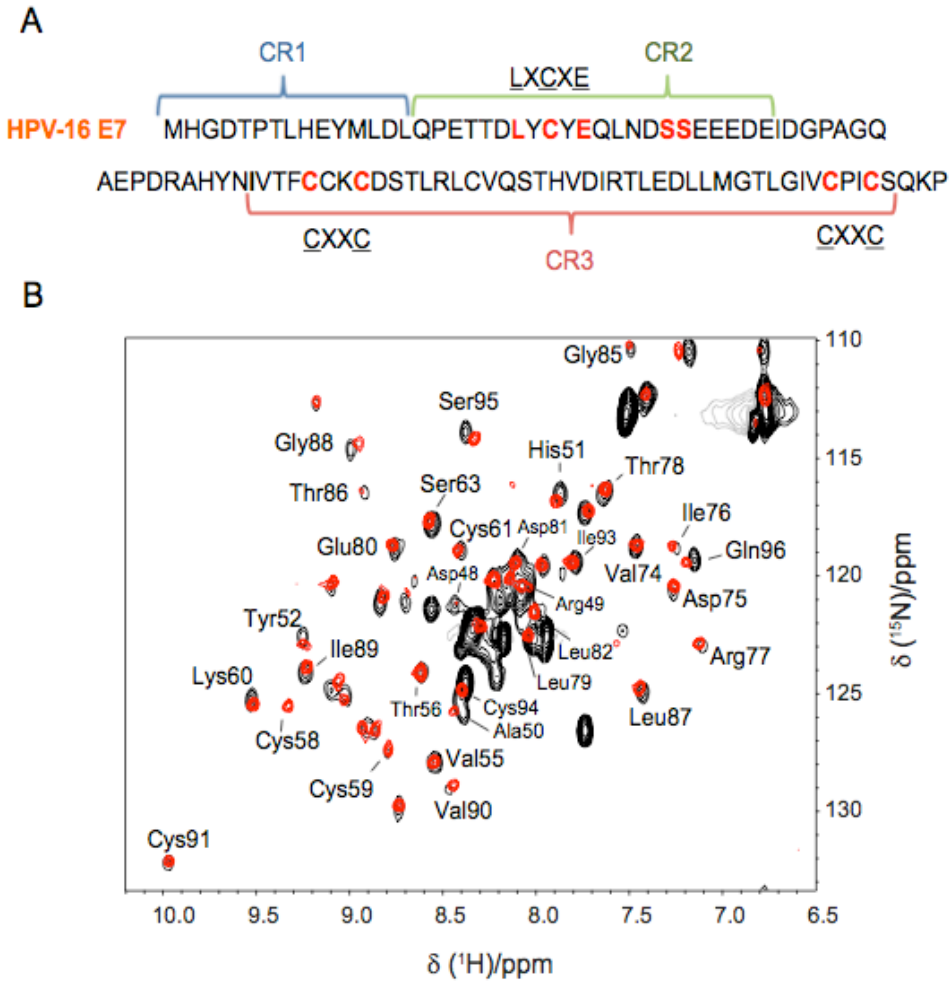


Figure 1. A) Amino acid sequence of HPV-16 E7, where specific motifs are highlighted in red: the LXCXE motif for pRb binding, the CKII phosphorylation site in CR2, and the zinc-binding motifs CXXC in CR3. B) ^{1}H-^{15}N SOFAST-HMQC spectra recorded for HPV-16 E7 (black) and for HPV-16 E7CR3 (red). The resulting assignment is reported in the figure, indicating the amino acid residue. The experiments were recorded at 11.7 T on samples 0.30 mM E7CR3 and 0.34 mM E7 in 50 mM HEPES, pH 8.5, 50 mM KCl, 10 μM $ZnCl_2$ and 5 mM DTT.

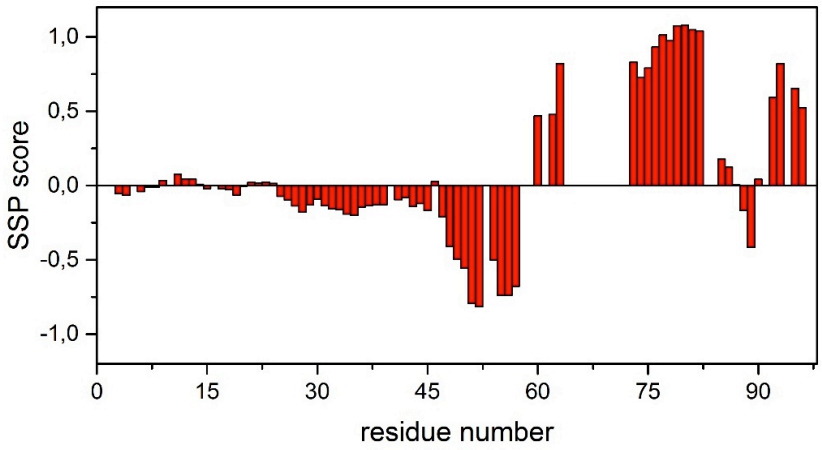


Figure 2. Secondary structural propensity (SSP)[35] score obtained by combining N, C', C^{α}, and C^{β} chemical shifts for the assigned resonances of HPV-16 E7 (data for the N-terminal half were taken from [56]).

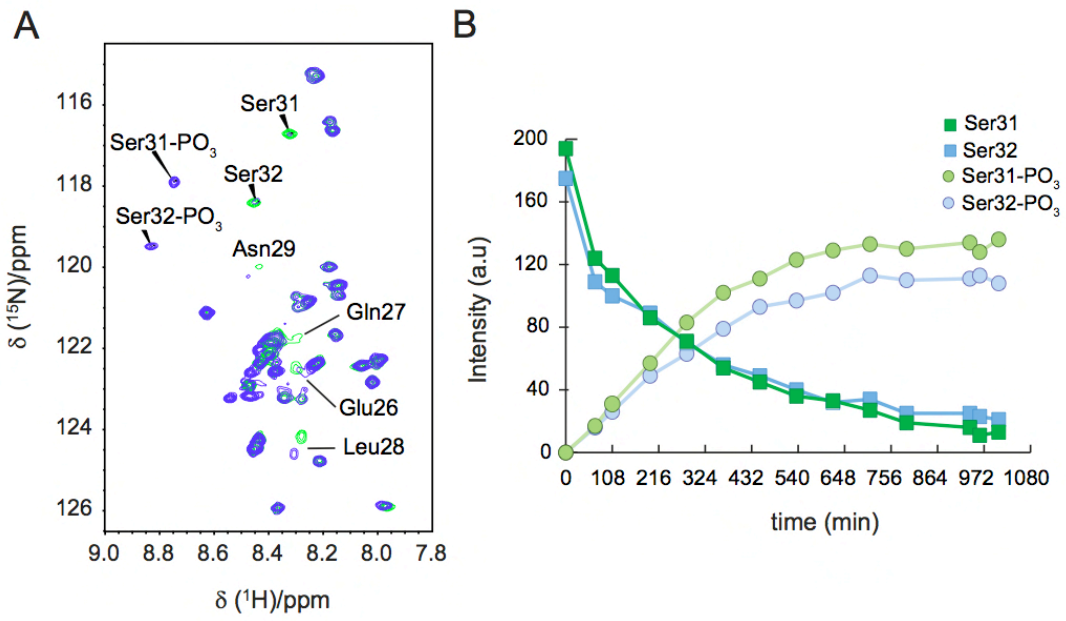


Figure 3. A) Superposition of 2D ^{1}H-^{15}N HSQC spectra of E7 before (green) and after (blue) phosphorylation (acquired after 414 min of reaction). B) The intensity changes of Ser31, Ser32 and their respective phosphorylated products were followed over time. Experiments were recorded at 303K at 16.4 T NMR spectrometers. Buffer conditions for the phosphorylation assay: 340 μM E7 in 20 mM HEPES pH7.5, 50 mM KCl, 10 mM DTT, 10 μM $ZnCl_2$; 800 μM Mg-ATP, 5000 UI CKII.

Table 1. Sequence specific assignment of H^N, N, C', C^α, and C^β in HPV-16 E7CR3. Proton resonances were calibrated with respect to the signal of 2,2-dimethylsilapentane-5-sulfonic acid (DSS). Carbon and nitrogen resonances were referenced indirectly to the ^1H standard using the conversion factor derived from the ratio of NMR frequencies [42]. The data have been deposited to the BMRB with accession number 26069.

Residue n°	A	H^N	N	C^α	C^β	C'
45	A			52.299	19.266	173.93
46	E	8.331	121.697	54.332		
47	P			63.1	31.956	176.417
48	D	8.455	120.689	54.573	41.041	175.901
49	R	8.110	120.219	55.214	32.632	174.792
50	A	8.494	125.49	50.817	21.219	174.727
51	H	7.932	116.60	55.878	32.508	175.293
52	Y	9.284	122.751	57.522	40.044	
53	N			52.804		174.270
54	I	9.075	124.900	60.075		174.773
55	V	8.566	127.639	61.837		175.267
56	T	8.674	123.831	58.428	69.480	170.94
57	F	8.287	119.959	55.373	41.769	175.153
58	C	9.369	125.214	59.125		178.164
59	C	8.853	127.143	59.73	27.542	174.048
60	K	9.578	125.272	57.611	33.806	177.301
61	C	8.471	118.714	59.161		176.082
62	D	8.197	119.847	56.931	40.706	175.255
63	S	8.639	117.413	60.242		
…						
73	H			59.407	30.220	177.145
74	V	7.497	118.474	65.485		178.249
75	D	7.308	120.287	57.253		177.978

76	I	7.309	118.536	60.820		176.710
77	R	7.172	122.696	58.272	29.211	178.547
78	T	7.681	116.151	66.399	67.980	176.382
79	L	8.090	122.277	58.333	41.617	177.846
80	E	8.812	118.524	60.672	29.647	179.315
81	D	8.148	119.128	57.633	39.993	180.132
82	L	8.068	121.238	58.096		
83	L					
84	M					176.454
85	G	7.566	110.055	45.333		174.716
86	T	8.985	116.231	61.913	68.962	173.881
87	L	7.513	124.51	55.903		174.519
88	G	9.014	114.181	44.083		171.823
89	I	9.292	123.761	59.982	24.191	174.210
90	V	8.493	128.626	60.883	33.804	175.721
91	C	10.009	131.874	58.092		
92	P			65.474	32.132	178.975
93	I	7.867	119.219	63.976	37.299	179.316
94	C	8.466	124.621	64.186	29.364	178.051
95	S	8.383	113.839	60.392	63.396	174.493
96	Q	7.248	119.203	55.797		
97	K	7.605	122.545	54.378		
98	P					

REFERENCES

1. Bernard, H.-U. *et al*. Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology* **401,** 70–79 (2010).

2. zur Hausen, H. Papillomavirus infections--a major cause of human cancers. *Biochim. Biophys. Acta* **1288,** F55-78 (1996).

3. de Villiers, E.-M., Fauquet, C., Broker, T. R., Bernard, H.-U. & zur Hausen, H. Classification of papillomaviruses. *Virology* **324,** 17–27 (2004).

4. zur Hausen, H. Papillomaviruses and cancer: from basic studies to clinical application. *Nat. Rev. Cancer* **2,** 342–350 (2002).

5. Münger, K. *et al*. Mechanisms of human papillomavirus-induced oncogenesis. *J. Virol*. **78,** 11451–11460 (2004).

6. Barbosa, M. S. The oncogenic role of human papillomavirus proteins. *Crit. Rev. Oncog*. **7,** 1–18 (1996).

7. Werness, B. A., Levine, A. J. & Howley, P. M. Association of human papillomavirus types 16 and 18 E6 proteins with p53. *Science* **248,** 76–79 (1990).

8. Dyson, N., Howley, P. M., Münger, K. & Harlow, E. The human papilloma virus-16 E7 oncoprotein is able to bind to the retinoblastoma gene product. *Science* **243,** 934–937 (1989).

9. Halbert, C. L., Demers, G. W. & Galloway, D. A. The E7 gene of human papillomavirus type 16 is sufficient for immortalization of human epithelial cells. *J. Virol*. **65,** 473–478 (1991).

10. Riley, R. R. *et al*. Dissection of human papillomavirus E6 and E7 function in transgenic mouse models of cervical carcinogenesis. *Cancer Res*. **63,** 4862–4871 (2003).

11. Phelps, W. C., Münger, K., Yee, C. L., Barnes, J. A. & Howley, P. M. Structure-function analysis of the human papillomavirus type 16 E7 oncoprotein. *J. Virol*. **66,** 2418–2427 (1992).

12. Calçada, E. O., Felli, I. C., Hošek, T. & Pierattelli, R. The Heterogeneous Structural Behavior of E7 from HPV16 Revealed by NMR Spectroscopy. *ChemBioChem* **14,** 1876–1882 (2013).

13. Ohlenschläger, O. *et al*. Solution structure of the partially folded high-risk human papilloma virus 45 oncoprotein E7. *Oncogene* **25,** 5953–5959 (2006).

14. Liu, X., Clements, A., Zhao, K. & Marmorstein, R. Structure of the human Papillomavirus E7 oncoprotein and its mechanism for inactivation of the retinoblastoma tumor suppressor. *J. Biol. Chem*. **281,** 578–586 (2006).

15. Todorovic, B. *et al*. Systematic Analysis of the Amino Acid Residues of Human Papillomavirus Type 16 E7 Conserved Region 3 Involved in Dimerization and Transformation. *J. Virol*. **85,** 10048–10057 (2011).

16. Davies, R., Hicks, R., Crook, T., Morris, J. & Vousden, K. Human papillomavirus type 16 E7 associates with a histone H1 kinase and with p107 through sequences necessary for transformation. *J. Virol*. **67,** 2521–2528 (1993).

17. McIntyre, M. C., Ruesch, M. N. & Laimins, L. A. Human papillomavirus E7 oncoproteins bind a single form of cyclin E in a complex with cdk2 and p107. *Virology* **215,** 73–82 (1996).

18. Funk, J. O. *et al*. Inhibition of CDK activity and PCNA-dependent DNA replication by p21 is blocked by interaction with the HPV-16 E7 oncoprotein. *Genes Dev*. **11,** 2090–2100 (1997).

19. Jones, D. L., Alani, R. M. & Münger, K. The human papillomavirus E7 oncoprotein can uncouple cellular differentiation and proliferation in human keratinocytes by abrogating p21Cip1-mediated inhibition of cdk2. *Genes Dev*. **11,** 2101–2111 (1997).

20. Massimi, P., Pim, D., Storey, A. & Banks, L. HPV-16 E7 and adenovirus E1a complex formation with TATA box binding protein is enhanced by casein kinase II phosphorylation. *Oncogene* **12,** 2325–2330 (1996).

21. Lee, J. O., Russo, A. A. & Pavletich, N. P. Structure of the retinoblastoma tumour-suppressor pocket domain bound to a peptide from HPV E7. *Nature* **391,** 859–865 (1998).

22. Chemes, L. B., Sánchez, I. E., Smal, C. & de Prat-Gay, G. Targeting mechanism of the retinoblastoma tumor suppressor by a prototypical viral oncoprotein. Structural modularity, intrinsic disorder and phosphorylation of human papillomavirus E7. *FEBS J*. **277,** 973–988 (2010).

23. Todorovic, B. *et al*. Conserved Region 3 of Human Papillomavirus 16 E7 Contributes to Deregulation of the Retinoblastoma Tumor Suppressor. *J. Virol*. **86,** 13313–13323 (2012).

24. Jansma, A. L. *et al*. The High-Risk HPV16 E7 Oncoprotein Mediates Interaction between the Transcriptional Coactivator CBP and the Retinoblastoma Protein pRb. *J. Mol. Biol*. **426,** 4030–4048 (2014).

25. Barbosa, M. S. *et al*. The region of the HPV E7 oncoprotein homologous to adenovirus E1a and Sv40 large T antigen contains separate domains for Rb binding and casein kinase II phosphorylation. *EMBO J*. **9,** 153–160 (1990).

26. Dick, F. A. & Dyson, N. J. Three Regions of the pRB Pocket Domain Affect Its Inactivation by Human Papillomavirus E7 Proteins. *J. Virol*. **76,** 6224–6234 (2002).

27. Singh, M., Krajewski, M., Mikolajka, A. & Holak, T. A. Molecular Determinants for the Complex Formation between the Retinoblastoma Protein and LXCXE Sequences. *J. Biol. Chem*. **280,** 37868–37876 (2005).

28. Jones, D. L., Thompson, D. A. & Münger, K. Destabilization of the RB tumor suppressor protein and stabilization of p53 contribute to HPV type 16 E7-induced apoptosis. *Virology* **239,** 97–107 (1997).

29. Genovese, N. J., Banerjee, N. S., Broker, T. R. & Chow, L. T. Casein Kinase II Motif-Dependent Phosphorylation of Human Papillomavirus E7 Protein Promotes p130 Degradation and S-Phase Induction in Differentiated Human Keratinocytes. *J. Virol*. **82,** 4862–4873 (2008).

30. Tamiola, K. & Mulder, F. A. A. Using NMR chemical shifts to calculate the propensity for structural order and disorder in proteins. *Biochem. Soc. Trans*. **40,** 1014–1020 (2012).

31. Massimi, P., Pim, D. & Banks, L. Human papillomavirus type 16 E7 binds to the conserved carboxy-terminal region of the TATA box binding protein and this contributes to E7 transforming activity. *J. Gen. Virol*. **78 (Pt 10),** 2607–2613 (1997).

32. Massimi, P. & Banks, L. Differential Phosphorylation of the HPV-16 E7 Oncoprotein during the Cell Cycle. *Virology* **276,** 388–394 (2000).

33. Firzlaff, J. M., Lüscher, B. & Eisenman, R. N. Negative charge at the casein kinase II phosphorylation site is important for transformation but not for Rb protein binding by the E7 protein of human papillomavirus type 16. *Proc. Natl. Acad. Sci. U. S. A.* **88,** 5187–5191 (1991).

34. Bienkiewicz, E. A. & Lumb, K. J. Random-coil chemical shifts of phosphorylated amino acids. *J. Biomol. NMR* **15,** 203–206 (1999).

35. Münger, K. *et al*. Complex formation of human papillomavirus E7 proteins with the retinoblastoma tumor suppressor gene product. *EMBO J.* **8,** 4099–4105 (1989).

36. Alonso, L. G. *et al*. The HPV16 E7 viral oncoprotein self-assembles into defined spherical oligomers. *Biochemistry (Mosc.)* **43,** 3310–3317 (2004).

37. Münger, K. *et al*. Complex formation of human papillomavirus E7 proteins with the retinoblastoma tumor suppressor gene product. *EMBO J.* **8,** 4099–4105 (1989).

38. Favier, A. & Brutscher, B. Recovering lost magnetization: polarization enhancement in biomolecular NMR. *J. Biomol. NMR* **49,** 9–15 (2011).

39. Lescop, E., Kern, T. & Brutscher, B. Guidelines for the use of band-selective radiofrequency pulses in hetero-nuclear NMR: example of longitudinal-relaxation-enhanced BEST-type 1H-15N correlation experiments. *J. Magn. Reson. San Diego Calif 1997* **203,** 190–198 (2010).

40. Palmer, A. G., Cavanagh, J., Wright, P. E. & Rance, M. Sensitivity improvement in proton-detected two-dimensional heteronuclear correlation NMR spectroscopy. *J. Magn. Reson. 1969* **93,** 151–170 (1991).

41. Kay, L., Keifer, P. & Saarinen, T. Pure absorption gradient enhanced heteronuclear single quantum correlation spectroscopy with improved sensitivity. *J. Am. Chem. Soc.* **114,** 10663–10665 (1992).

42. Schleucher, J. *et al*. A general enhancement scheme in heteronuclear multidimensional NMR employing pulsed field gradients. *J. Biomol. NMR* **4,** 301–306 (1994).

43. Lescop, E., Schanda, P. & Brutscher, B. A set of BEST triple-resonance experiments for time-optimized protein resonance assignment. *J. Magn. Reson. San Diego Calif 1997* **187,** 163–169 (2007).

44. Salzmann, M., Pervushin, K., Wider, G., Senn, H. & Wüthrich, K. TROSY in triple-resonance experiments: new perspectives for sequential NMR assignment of large proteins. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 13585–13590 (1998).

45. Grzesiek, S. & Bax, A. Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance NMR. *J. Am. Chem. Soc.* **114,** 6291–6293 (1992).

46. Kay, L. E., Xu, G. Y., Singer, A. U., Muhandiram, D. R. & Formankay, J. D. A Gradient-Enhanced HCCH-TOCSY Experiment for Recording Side-Chain 1H and 13C Correlations in H2O Samples of Proteins. *J. Magn. Reson. B* **101,** 333–337 (1993).

47. Geen, H. & Freeman, R. Band-selective radiofrequency pulses. *J. Magn. Reson. 1969* **93,** 93–141 (1991).

48. Vranken, W. F. *et al.* The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins* **59,** 687–696 (2005).

49. Keller, R. *The computer aided resonance assignment tutorial*. (CANTINA, 2004).

50. Schanda, P. & Brutscher, B. Very Fast Two-Dimensional NMR Spectroscopy for Real-Time Investigation of Dynamic Events in Proteins on the Time Scale of Seconds. *J. Am. Chem. Soc.* **127,** 8014–8015 (2005).

51. Schanda, P., Kupce, E. & Brutscher, B. SOFAST-HMQC experiments for recording two-dimensional heteronuclear correlation spectra of proteins within a few seconds. *J. Biomol. NMR* **33,** 199–211 (2005).

52. Grzesiek, S. & Bax, A. The importance of not saturating water in protein NMR. Application to sensitivity enhancement and NOE measurements. *J. Am. Chem. Soc.* **115,** 12593–12594 (1993).

Section 3.3

The structure and dynamics of HPV-16 E7 described by NMR-guided metadynamics

Magnetic Resonance Center (CERM) and Department of Chemistry "Ugo Schiff", University of Florence, Italy

Department of Chemistry, University of Cambridge, Cambridge, United Kingdom

In preparation

INTRODUCTION

The high-risk HPV-16 E7 protein is one of the most striking examples of proteins that possess partially folded secondary structure and, in contrast to common globular proteins, exists as a dynamic ensemble of flexible conformers. Such intrinsically disordered proteins (IDPs) fail to form a stable structure and their backbone dihedral angles vary significantly over time with no specific equilibrium values [1].

The polypeptide chain of HPV-16 E7 consists of 98 amino acids and is usually subdivided into three conserved regions called CR1 (a.a. 1-15), CR2 (a.a 16-40), and CR3 (a.a. 41-98). Evidence suggests that the CR3 region adopts a folded three-dimensional arrangement and is actively involved in protein dimerization[2–4]. By contrast, CR1 and CR2 have properties of intrinsically disordered regions and their flexibility is thought to mediate interactions that contribute to HPV-mediated oncogenesis [5].

Many attempts have been made to fully characterise the structure of E7 either by simulation or experimentally, but it still remains an important open issue to be settled. We thus decided to apply an NMR-guided metadynamics approach to describe the structure and the dynamics of E7.

RESULTS AND DISCUSSION

In order to determine the conformational properties of E7, we performed molecular dynamics simulations using replica-averaged metadynamics (RAM). This technique combines the sampling efficiency of metadynamics with experimentally restrained force fields. All the simulations were run using the bias-exchange sampling method and included chemical shifts and residual dipolar couplings as ensemble-averaged structural restraints [6]. This allowed us to define a system-dependent potential acting on the molecule and to increase the accuracy of the conformational space sampling. Several independent simulations were carried out on four different replicas of the system and the trajectory of each replica was biased by a different collective variable. As recently demonstrated by Cavalli and co-workers [7], the ensemble of structures determined with this method agrees with the maximum entropy principle and is the best possible result that satisfies the experimental data.

The first and the second collective variables (CVs) used in the simulations were chosen to bias the backbone dihedral angles ϕ and ψ of all residues in the protein. The third collective variable acted on the side-chain dihedral angle χ_1 of all the heavy side-chain residues (His, Tyr, Met, Arg, Phe, and Lys). The fourth collective variable was used to calculate the electrostatic energy between the two monomeric subunits using an approximated form of the Debye-Hückel equation [7].

The dynamic evolution of E7 during a period of 250 ns per replica (for a total simulation time of 1 μs) was simulated using the RAM approach. The first 200 ns were needed for the system to reach convergence while the additional 50 ns were spent to thoroughly sample the free energy landscape of the protein. Figure 1 reports the free energy as a function of the four collective variables. Error bars correspond to standard deviations ≤ 3.5 kJ/mol.

The inclusion of NMR data (N^H, C^α, C^β and C' chemical shifts and RDCs) in the simulations made a fundamental difference in terms of convergence time and model accuracy. The structural restraints forced the protein to assume conformations consistent with the measured data, which resulted in an increased average value for the root-mean-square deviations of atomic positions (12.9 ± 0.9 Å for NMR-guided simulations, 7.8 ± 0.94 Å for unrestrained simulations). The conformational ensemble calculated with RAM, therefore, was substantially different from that predicted by standard metadynamics and the population of structures was characterised by much more opened geometries.

In order to further validate RAM simulations, we used the resulting pools of structures to back-calculate NMR chemical shifts with the SPARTA+ algorithm [8]. As SPARTA+ is based on different principles than CamShift [9], which was here used during replica-averaged metadynamics, the resulting agreement provided an independent validation of the quality of the ensemble. Separate backbone and C^β chemical shifts were predicted for both restrained and un- restrained simulations and the ensemble-averaged values were given as Boltzmann- weighted averages over all the conformers. The comparison of our simulations with the experimental data is illustrated in Figure 2. Unrestrained metadynamics showed almost no correlation with experimental chemical shifts, with squared correlation coefficients at most equal to 0.0671. By contrast, the

incorporation of NMR data into the force field led to a significant increase in R^2 values and to a more accurate sampling of the conformational space. All the correlation coefficients reached values close to 1, with the exception of amide hydrogens whose R^2 is 0.3588. This discrepancy is common to all the major chemical shift prediction programs and is due to the sensitivity of proton chemical shifts to chemical exchange, solvent exposure, pH, ionic strength, as well as to structural effects like hydrogen bonding, electric fields, and ring currents [10].

In the same fashion, measured RDC values were used as criteria for the validation of the ensembles generated with both RAM and unrestrained simulations (Figure 2). Since RDCs provide information about how bond vectors between nuclei are aligned relative to each other, the use of such data in protein molecular dynamics is an effective way to simulate the folding of a polypeptide into secondary and tertiary structural elements. The agreement between experimental and calculated RDC values was assessed by using the quality factor (Q) as proposed by Cornilescu and co-workers [11]. The Q factor obtained from the RAM ensemble was 0.2715, whereas the corresponding value for the unrestrained ensemble was 0.8465. The lower value of the former quality factor demonstrates that the corresponding ensemble was in good agreement with the experimental residual dipolar couplings and that the RAM method was able to capture the dynamics of E7 better than standard metadynamics.

Traditionally, protein folding is illustrated by 1-D profiles such as root-mean-square deviations, energy, radius of gyration and inter-residue distances. The hyper-dimensional nature of a protein's free energy landscape makes none of these monodimensional representations adequate to properly describe the thermodynamics of folding processes. The opportunity of shifting the attention from the enormous number of atomic positions to a restricted set of degrees of freedom (often referred to as collective variables, CVs) greatly simplifies the analysis of protein dynamics and permits to condense all the information in a few basic parameters. The combination of any pair of collective variables with the free energy coordinate ΔF results in a three-dimensional energy surface (or landscape) which maps every possible conformation of the protein to its corresponding energy level. From the overall shape of such surfaces, the folding rate, the flexibility, and the native structure of the protein

can be argued. Rapidly foldable proteins have smooth "funnel-shaped" energy surfaces with a single basin of attraction corresponding to the native state. Disordered proteins have rugged surfaces with multiple local minima separated by small energy barriers. Halfway between the two classes are partially folded proteins whose free energy landscape has a funnel-like shape and is characterised by many local minima [12].

The three-dimensional (ϕ, ψ, ΔF) free energy landscape of E7 and its projection onto the plane of CVs are given in Figure 3. As in the case of proteins having one or more IDRs, the calculated energy surface displays two major basins surrounded by a series of shallower local minima. Each well corresponds to a microstate of the system and each microstate is populated by a multitude of isoenergetic conformers. The most populated basin (55% population, labelled A in the figure) corresponds to ϕ = 84.8°, ψ = 93.9°, χ_1 = 19.7°, and E_{DH} = 2.1 kJ/mol. The second most populated basin (7% population, labelled B in the figure) is characterised by ϕ = 82.9°, ψ = 85.4°, χ_1 = 15.1°, and E_{DH} = 0.9 kJ/mol. A remaining 35% of the ensemble is distributed over the portion of the energy surface ranging between 7.5 and 15 kJ/mol and only 3% of the structures populates states with energy > 15 kJ/mol. The energy barrier that separates neighbouring local minima is generally small compared with *RT* and transitions between different conformations of the protein can easily occur under physiological conditions. This is consistent with the idea that structures of different microstates are in fast exchange on the NMR time scale and every single conformer makes its contribution to the observed chemical shifts and RDCs. This is supported by the fact that neglecting even the least probable structures in the ensemble led to an incorrect back-calculation of the NMR observables and sensibly decreased the correlation between theoretical and experimental data.

Even if E7 should always be described in terms of an ensemble of states, analysis of the most stable conformations can give an insight into the general features of the protein's structure. An example of such conformations is given in Figure 4 (left panel). In accordance with confirmed experimental evidences [13,14], all RAM simulations produced a homodimer model whose subunits interacted predominantly through the hydrophobic residues in the C-terminal domain. Analysis of secondary structural elements during the simulations confirmed that the N-terminus is characterised by a

high flexibility and a marked disorder content (Figure 4, right panel). Most of its residues form a coiled structure with a probability greater than 96% and only a few of them show a limited propensity for folding into structured elements. As already noted, residues 10 to 15 populate a conformationally fluctuating α-helix element with a probability of ~ 20%. This suggests that the second half of CR1 is subject to interresidue forces, which stabilise the structure and lead to a conformational equilibrium involving coil-helix transitions.

The LXCXE motif in CR2 demonstrated a high propensity for remaining in an unfolded state, having only a 2% of probability of populating a short α-helix element. This section of CR2 is known to have binding affinity for the pRb protein [15] and its disorder content could be partially responsible for modulating the interaction of E7 with its substrates. Our finding contradicts other studies [16,17] which presented the LXCXE sequence as a statically folded motif. The hypothesis we present is supported by the evidence that a highly disordered nature of E7 is always required for the development of high-grade lesions and cancer [18]. In fact, while E7 proteins of high-risk HPV types 11 and 18 have a prominent coil-like nature, HPV-1 E7 is less disordered and is not responsible for the onset of cervical cancer. The intrinsically disordered character of CR1 and CR2 may thus have two different functions. The first function is providing to E7 enough mobility for sampling a wide area in the search for its partners. The second is related to the possibility of a binding-induced folding of residues 10-15. After the highly complementary interaction of LXCXE with pRb, in fact, vicinal amino acids could be affected by the new arrangement of the domain and could undergo a conformational transition to stabilise the structure and the binding site.

Inspection of the dynamics of the N-terminus also revealed that CR1 tends to bend back on itself and sometimes to associate with the same region of the other subunit. The interaction is dominated by van der Waals forces and is responsible for the formation of small unstructured clusters. In fact, hydrophobic amino acids present in CR1 and CR2 tend to repel the aqueous environment and to cluster together, stabilising the protein structure. The phenol group of Tyr, in particular, is prone to interact with other aromatic ring systems and to create a hydrophobic moiety in the middle of the otherwise highly hydrophilic coil. This decreases the free energy content

of the structure and helps stabilising apolar side-chains of neighbouring residues.

The CR3 region in the C-terminal domain is characterised by a folded structure containing a major right-handed α-helix (a.a. 72-83), an antiparallel β-sheet (a.a. 50-69) and a minor right-handed α-helix formed by residues 91 to 95. The zinc finger domain is located in between these folded elements and comprises two CXXC motifs. The first CXXC motif is formed by residues 58 to 61 and is always part of a small unstructured coil. The second motif lies between Cys-91 and Cys-94 and has a probability of ~ 20% of populating the minor α-helix of CR3. The structure is held together by the donor-acceptor interaction between the sulphur atoms of cysteines and the Zn^{2+} ion.

Analysis of the interdomain linkers within CR3 revealed that the segments corresponding to a.a. 57-62 and 70-78 continuously fluctuate in space and generally adopt a random coil conformation. The disordered behaviour of inter-domain linkers is a common feature of IDPs [19] and is one of the main contribution to the overall motion of partially structured proteins. Interestingly, the linker between the two β-strands contains one of the two CXXC motifs, a fact that seems to be inconsistent with the rigidity required by the zinc finger domain. Looking at the *relative* position of the four Cys, however, it is clear that sulphur atoms always maintain their positions at the vertices of the tetrahedron centred on the zinc (Figure 3.12(b)). Therefore, the whole zinc finger domain behaves as a rigid structure able to adjust itself in response to major structural modifications of the N-terminal regions but not to get deformed. It is only allowed to translate in space to follow the wiggling motion of the inter-domain linker.

The partial plasticity of CR3 confirms earlier suggestions that this region plays a role in the binding mechanism of E7 with the pRb protein. As reported by Liu and co-workers [20], the two segments 57-62 and 70-78 are thought to be involved in the formation of the E7-pRb complex thanks to their marked disorder content. After the formation of the stable E7-pRb complex, the two coiled segments tend to make contact with the bound pRb and establish an interaction that inhibits the regulatory function of the substrate. The low affinity and high specificity of such interaction resembles that of many IDP binding mechanisms and suggest that the localised disorder of CR3 could be essential for E7's biological functions.

CONCLUSIONS

In this work we have demonstrated that NMR-driven replica-averaged metadynamics is able to describe the conformational space of the HPV-16 E7. Bias-exchange metadynamics enhances the sampling of the conformational space whereas NMR restraints continuously modify the force field to increase the consistency of simulations with experimental measurements. In this way, we have been able to reconstruct the free energy landscape of the protein and to obtain a conformational ensemble that represents the dynamics of its domains.

The high-quality three-dimensional models of E7 obtained in this work can be used for designing new drugs with desired anti-HPV properties. *In vitro* and *in vivo* experimentations are needed for a better understanding of the exact interaction mechanism of E7 with its partners, but the validation of our structural model could provide great opportunities to decrease subclinical infections and to expand the pharmaceutical arsenal against cancer and cervical lesions.

ACKNOWLEDGEMENTS

This work has been supported by the European Commission Projects Marie Curie ITN IDPbyNMR (Contract No. 264257), BioNMR (Contract No. 261863) and INSTRUCT (Contract No. 211252). MON is grateful to the program “Science without borders” of the Brazilian Ministry of Science and Technology (CNPq) for financial support. MLP was a recipient of an Erasmus placement fellowship from the University of Florence (Italy).

FIGURES

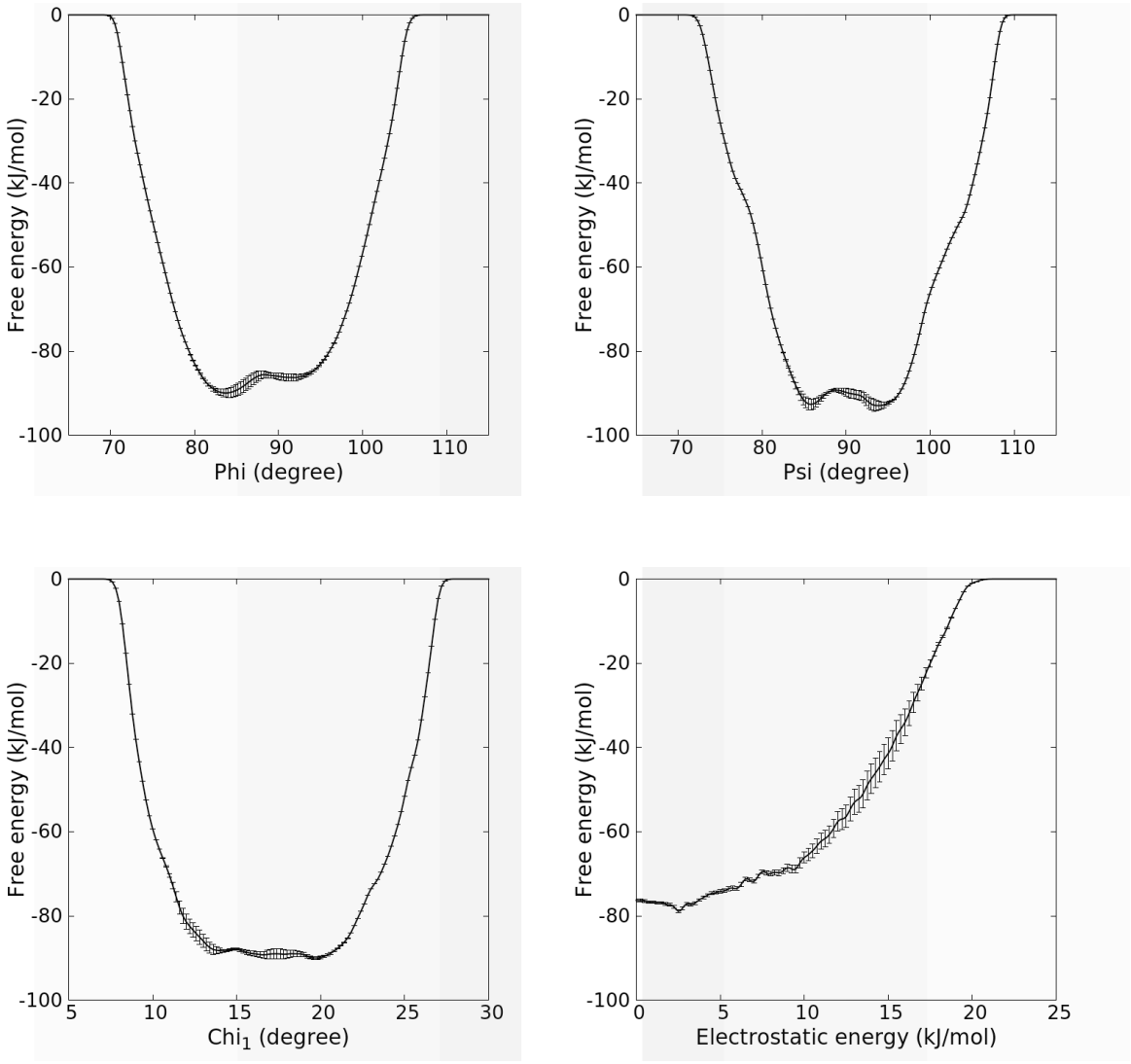
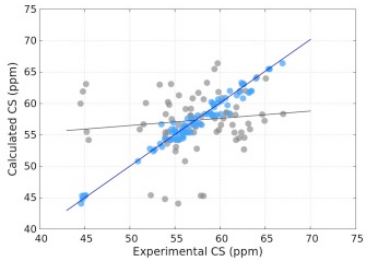
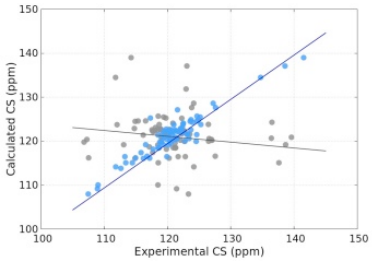


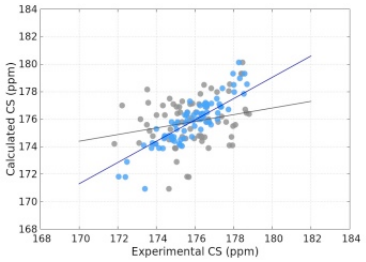
Figure 1. Free energy landscapes as a function the four used collective variables. The projected surfaces are averaged over the last 50 ns of the simulations and their standard deviations are reported as error bars.



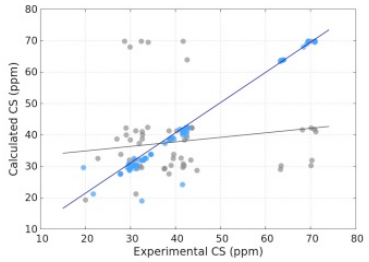
C^{α} chemical shifts. Unrestrained simulations: $R^2 = 0.0116$, slope = 0.11. NMR- guided simulations: $R^2 = 0.9604$, slope = 1.00.



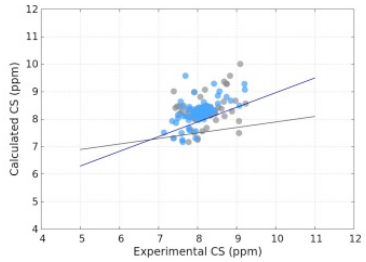
N chemical shifts. Unrestrained simulations: $R^2 = 0.0118$, slope = –0.13. NMR-guided simulations: $R^2 = 0.8768$, slope = 1.00.



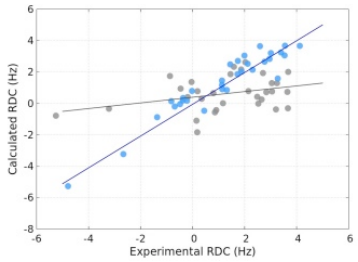
C' chemical shifts. Unrestrained simulations: $R^2 = 0.0671$, slope = 0.24. NMR-guided simulations: $R^2 = 0.8115$, slope = 0.78.



C^{β} chemical shifts. Unrestrained simulations: $R^2 = 0.0124$, slope = 0.14. NMR-guided simulations: $R^2 = 0.9477$, slope = 0.96.



H^{N} chemical shifts. Unrestrained simulations: $R^2 = 0.0023$, slope = 0.20. NMR-guided simulations: $R^2 = 0.3588$, slope = 0.53.



RDCs. Unrestrained simulation: Q = 0.8465, slope = 0.18. NMR-guided simulation: Q = 0.2715, slope = 0.90.

Figure 2. Scatter diagrams showing correlation between experimental and back-calculated NMR chemical shifts and RDCs. Unrestrained metadynamics results are shown in grey and NMR-guided RAM results are shown in blue.

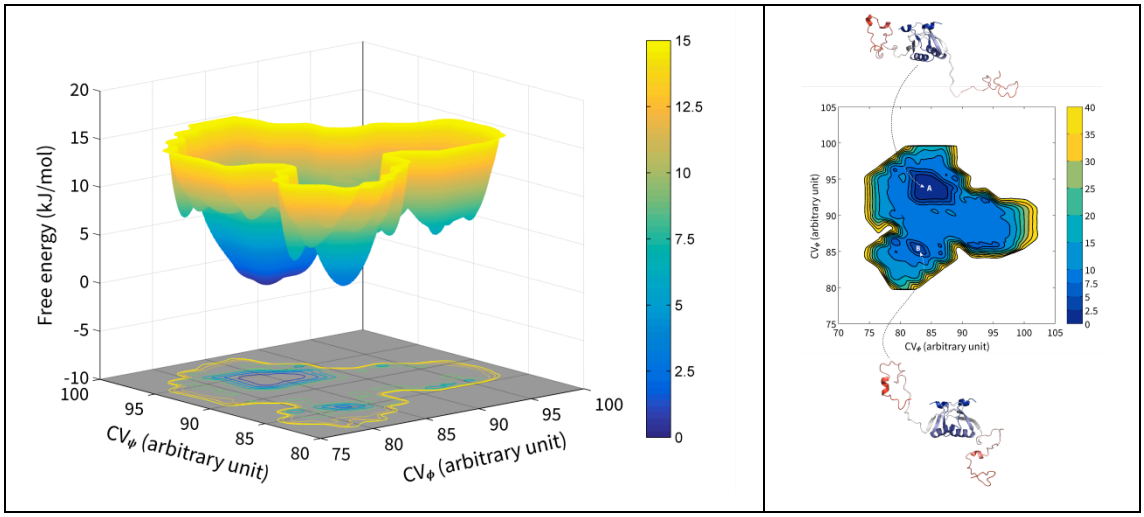


Figure 3. Left Panel: Three-dimensional free energy landscape as a function of the collective variables for the ϕ and ψ dihedral angles. Right panel: Characterisation of the microstates of the E7 protein. The free energy landscape (in kJ/mol) is reported as a projection on the plane of the two collective variables ϕ and ψ. A representative structure is shown for the two microstates with a free energy lower than 5 kJ/mol.

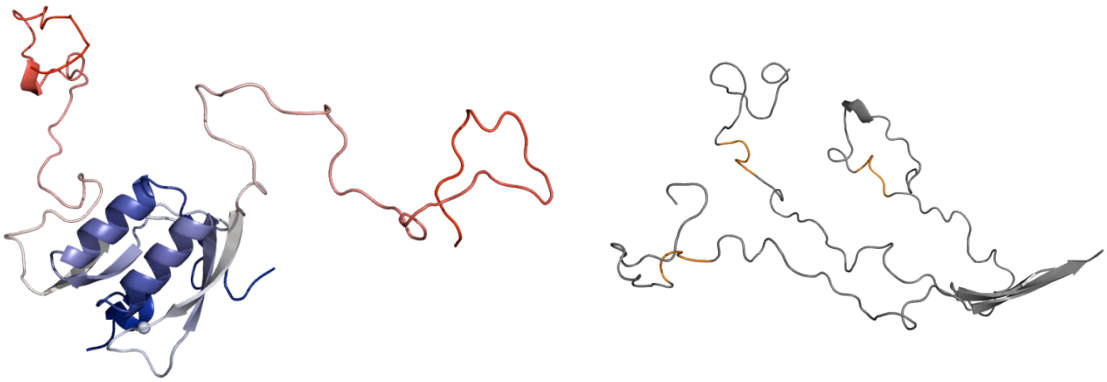


Figure 4. Left panel**:** Representative structure of a minimum-energy conformation of HPV-16 E7. The structure is coloured from the N-terminal in red to the C-terminal in blue and the two zinc ions are shown as spheres. Right panel: Ribbon diagram of three representative conformations of the CR1 and CR2 regions. The LXCXE motif of each structure is coloured in orange. All the structures illustrated in the text are visualised using PyMOL.

REFERENCES

1. Uversky, V. N. & Dunker, A. K. Understanding protein non-folding. *Biochim. Biophys. Acta BBA - Proteins Proteomics* **1804,** 1231–1264 (2010).

2. Calçada, E. O., Felli, I. C., Hošek, T. & Pierattelli, R. The Heterogeneous Structural Behavior of E7 from HPV16 Revealed by NMR Spectroscopy. *ChemBioChem* **14,** 1876–1882 (2013).

3. Todorovic, B. *et al*. Systematic Analysis of the Amino Acid Residues of Human Papillomavirus Type 16 E7 Conserved Region 3 Involved in Dimerization and Transformation. *J. Virol*. **85,** 10048–10057 (2011).

4. Alonso, L. G. *et al*. The HPV16 E7 viral oncoprotein self-assembles into defined spherical oligomers. *Biochemistry (Mosc.)* **43,** 3310–3317 (2004).

5. García-Alai, M. M., Alonso, L. G. & de Prat-Gay, G. The N-Terminal Module of HPV16 E7 Is an Intrinsically Disordered Domain That Confers Conformational and Recognition Plasticity to the Oncoprotein. *Biochemistry (Mosc.)* **46,** 10405–10412 (2007).

6. Camilloni, C., Cavalli, A. & Vendruscolo, M. Replica-Averaged Metadynamics. *J. Chem. Theory Comput*. **9,** 5610–5617 (2013).

7. Cavalli, A., Camilloni, C. & Vendruscolo, M. Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *J. Chem. Phys*. **138,** 94112 (2013).

8. Shen, Y. & Bax, A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR* **48,** 13–22 (2010).

9. Kohlhoff, K. J., Robustelli, P., Cavalli, A., Salvatella, X. & Vendruscolo, M. Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. *J. Am. Chem. Soc*. **131,** 13894–13895 (2009).

10. Mulder, F. A. A. & Filatov, M. NMR chemical shift data and ab initio shielding calculations: emerging tools for protein structure determination. *Chem. Soc. Rev*. **39,** 578–590 (2010).

11. Cornilescu, G., Marquardt, J. L., Ottiger, M. & Bax, A. Validation of Protein Structure from Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase. *J. Am. Chem. Soc.* **120,** 6836–6837 (1998).

12. Burger, V., Gurry, T. & Stultz, C. Intrinsically Disordered Proteins: Where Computation Meets Experiment. *Polymers* **6,** 2684–2719 (2014).

13. Clemens, K. E., Brent, R., Gyuris, J. & MüNger, K. Dimerization of the Human Papillomavirus E7 Oncoproteinin Vivo. *Virology* **214,** 289–293 (1995).

14. Ohlenschläger, O. *et al.* Solution structure of the partially folded high-risk human papilloma virus 45 oncoprotein E7. *Oncogene* **25,** 5953–5959 (2006).

15. Dyson, N., Howley, P. M., Münger, K. & Harlow, E. The human papilloma virus-16 E7 oncoprotein is able to bind to the retinoblastoma gene product. *Science* **243,** 934–937 (1989).

16. Lee, J. O., Russo, A. A. & Pavletich, N. P. Structure of the retinoblastoma tumour-suppressor pocket domain bound to a peptide from HPV E7. *Nature* **391,** 859–865 (1998).

17. Nicolau-Junior, N. & Giuliatti, S. Modeling and molecular dynamics of the intrinsically disordered e7 proteins from high- and low-risk types of human papillomavirus. *J. Mol. Model.* **19,** 4025–4037 (2013).

18. Sang, B. C. & Barbosa, M. S. Single amino acid substitutions in 'low-risk' human papillomavirus (HPV) type 6 E7 protein enhance features characteristic of the 'high-risk' HPV E7 oncoproteins. *Proc. Natl. Acad. Sci. U. S. A.* **89,** 8063–8067 (1992).

19. Kukic, P., Camilloni, C., Cavalli, A. & Vendruscolo, M. Determination of the individual roles of the linker residues in the interdomain motions of calmodulin using NMR chemical shifts. *J. Mol. Biol.* **426,** 1826–1838 (2014).

20. Liu, X., Clements, A., Zhao, K. & Marmorstein, R. Structure of the human Papillomavirus E7 oncoprotein and its mechanism for inactivation of the retinoblastoma tumor suppressor. *J. Biol. Chem.* **281,** 578–586 (2006).

Supplementary Material

Protein expression and purification

For the expression of the wild type HPV-16 E7 (amino acids 1-98) the recombinant vector pET20-E7-opt [1] containing the gene for expression of E7 protein (1-98-His_6), was mutated by the insertion of a stop codon (TAG) between the codon for Pro 98 and the codons for the His_6 tag, generating the pET20-E7-WT. Thus, that plasmid was used for transformation of *E. coli* BL21 pLys (Stratagene®) cells.

After expression tests, one colony was selected for the growing in LB medium with ampicillin (0.1 mg/mL) and chloramphenicol (0.34 mg/mL). Finally, the cells were cultured in 1 L of M9 medium at 37 °C containing $(^{15}NH_4)_2SO_4$, 0.1 mg/mL ampicillin, 0.34 mg/mL chloramphenicol and 100 μM $ZnCl_2$. When the culture reached OD_{600} 0.7 protein expression was induced by the addition of IPTG (isopropyl β-D-1-thiogalactopyranoside) during 5h at 37 ºC. After that the cells were harvested at 4.500 rpm for 20 min at 4°C. The supernatant was discarded and the cells were re-suspended in 25 mL of degassed buffer A (20 mM TRIS buffer at pH 8.5, 10 mM DTT supplemented with Roche complete protease inhibitors). After this the bacterial cells were disrupted by sonication (20 kHz sonication frequency and 40% of amplitude) with cycles of 3s and 10s of delay in anaerobic conditions. Lysed cells were then centrifuged at 165000 *g*. The supernatant containing E7 protein was loaded on a HiTrap QFF 10 mL ionic exchange column pre-equilibrated with buffer A. The purification step was followed with buffer A in a gradient with buffer B (buffer A + 0.5 M NaCl). E7 protein was eluted with 0.5 M NaCl (100% buffer B). Fractions containing the protein were mixed, concentrated to 4 mL and injected in a Superdex 75 16/100 size exclusion chromatography column pre-equilibrated with the final degassed buffer (10 mM HEPES pH 7.5, 50 mM KCl, 10 mM DTT 10 μM $ZnCl_2$). After analysis by SDS-PAGE 17%, fractions containing the pure protein were mixed and concentrated to 500 μL for NMR analysis. 50 μL of D_2O was added to each NMR sample for the lock signal.

NMR Restraints

Two system-dependent energy terms were included in the force field in order to bias the potential acting on the molecule and to better reproduce the local rearrangement of the protein. The two energy terms were defined as:

$$E_{CS} = \sum_{i=1}^{n} \sum_{j=1}^{6} E_{ij}(\delta_{ij}^{calc} - \delta_{ij}^{exp})$$

$$E_{RDC} = \alpha \sum_{i} \left(D_i^{exp} - D_i^{calc}\right)^2,$$

and relied on nuclear magnetic resonance (NMR) chemical shifts and residual dipolar couplings as structural restraints. Their function was to modify the force field during the simulations and to increase the consistency of the sampling with experimental measurements. In this way the RAM method was able to explore low-probability regions of the free energy landscape and to allow the free diffusion of the system along particular reaction coordinates (collective variables).

Sequence-specific assignment of H^N, N, C', C^α , and C^β resonances was achieved using a series of ^{1}H- and ^{13}C-detected NMR experiments at 298 K [2,3]. In the first case, measurements were carried out on a18.8 T Bruker Avance 800 operating at 800.13 MHz for ^{1}H; in the second case we used a 16.4 T Bruker Avance 700, operating at 176.03 MHz for ^{13}C. The assignments of the HPV-16 E7 protein have been reported by Calçada and co-workers [1] and are available in the BMRB database (Accession number 19442).

One-bond N-H RDCs ($^1D_{NH}$) were measured with a 2D ^{1}H-^{15}N HSQC-IPAP experiment in the aligned liquid crystalline media *n*-octyl-penta(ethylene glycol)/octanol (C8E5) [4,5]. All spectra were recorded on a Bruker Avance III spectrometer operating at 950 MHz for ^{1}H and at 15 Hz for ^{2}H. Experimental values are shown in the following Table (S1).

Table S1 Residual dipolar couplings (Hz) for HPV-16 E7 at 293 K.

Residue	$^1D_{NH}$	**Residue**	$^1D_{NH}$
4	2.150	29	-0.040
5	-0.460	30	2.710
8	-5.260	31	0.190
11	-3.220	32	0.350
16	0.150	33	0.800
18	0.860	34	3.210
19	3.060	35	3.670
20	1.870	36	2.030
21	0.910	37	3.050
22	0.180	38	2.840
23	-0.190	39	3.680
25	1.170	40	2.660
26	3.250	42	2.490
27	1.590	43	2.530
28	1.460	44	3.650

RAM calculations

The computational modelling of HPV-16 E7 was performed using replica-average metadynamics (RAM). This technique involves a combination of replica exchange [6] and metadynamics [7] and is today one of the most reliable methods for reconstructing the free energy of complex biological systems. All the simulations presented in this work were performed using GROMACS [8] implemented with PLUMED2 [9] and Almost [10,11].

The amino acid sequence was retrieved from the UniProt database (code P03129) and the first guess of the protein's three-dimensional structure was obtained by comparative modelling using the Rosetta server. The carboxy-terminal half of the protein was built using the CR3 region of HPV-45 E7 (PDB 2EWL) as a template; CR1 and CR2 were built using a *de novo* prediction protocol. Both the monomeric and dimeric forms of E7 were considered and different symmetries of the dimer were attempted. The top-ranking model is shown in Figure S1. The structure appeared to us too compact and lacking some of the features typical of IDPs/IDRs. The reason of such anomaly lies in the protocol Rosetta uses to assign the ranking score to the resulting structural models. In fact, Rosetta's scoring function is based on the energy of the polypeptide chain while the experimentally observed native state is at a minimum of the free energy, rather than the energy [12]. The neglect of the missing configurational entropy contribution to the free energy is irrelevant for folded states but is an important shortcoming in structure predictions of disordered proteins. The quality of subsequent MD simulations, however, was not affected by the compactness of the starting structure and the resulting equilibrium ensemble was populated by conformations significantly more disordered.

Once the comparative modelling was completed, the protein was solvated with 25,000 molecules of water in a 800 nm^3 dodecahedron box, and 32 Na^+ ions were added to neutralize excess charges. AMBER ff99SB* [13,14] was chosen as the force field acting on the molecule and solvent was explicitly accounted with the TIP3P water model [15]. The system was evolved with a time step of 2.0 fs by constraining the fast bonded modes to their equilibrium value with the LINCS algorithm [8]. Temperature

was kept constant at 298 K by the velocity-rescaling thermostat so no preliminary heating period was needed. Van der Waals and Coulomb interactions were truncated at the cut-off distance of 1.2 Å and 0.9 Å, respectively. The two zinc fingers were taken into account by coordinating each Zn^{2+} ion to four CYM residues, that is to four cysteines having deprotonated sulphur atoms s in place of the normal –SH groups. The mobility of zinc ions was reduced by an upper wall of 2.2 Å in order to avoid their departure from the zinc finger domains.

A total of six RAM simulations were run: two were unrestrained simulations, two were CS-guided simulations and the other two were simultaneously CS- and RDC-guided simulations. In each run four replicas of the system were considered, one per collective variable. Every simulation lasted for 250 ns, 200 ns of which were required to reach convergence and the successive 50 ns were devoted to free energy estimation.

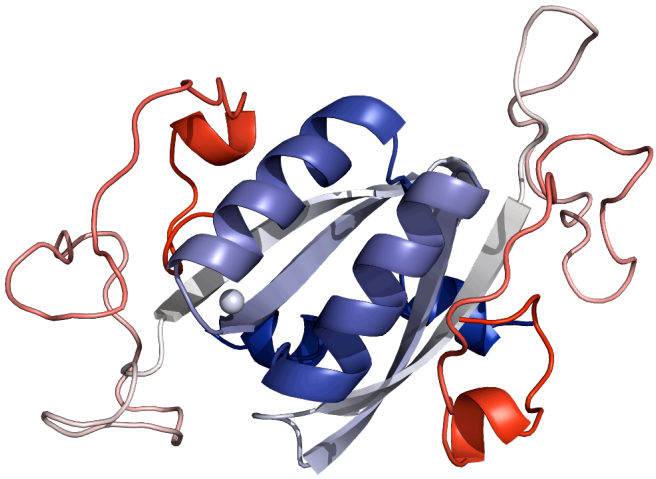


Figure S1. Comparative modelling of the dimeric structure of E7.

Collective Variables

The choice of collective variables (CVs) is a very important issue in metadynamics simulations since the evolution of the systems, and hence the time amount required to reach convergence, depends directly on their definition. In order to explore exhaustively the conformational space of E7, four collective variables related to specific molecular geometric properties were used. The first three collective variables were defined by the quantity 'alpha-beta similarity' that measured the

departure of the instantaneous values of a set of torsional angles from a set of reference values. The general formula is:

$$S = \frac{1}{2}\sum_i [1 + \cos(\theta_i - \theta_i^{\text{ref}})].$$

In our specific case, we monitored three different torsion angles (ϕ, ψ, χ_1). The first and second dihedrals defined the pair of Ramachandran angles (ϕ, ψ) and were thus for the determination of the structural elements formed by the backbone chain. The third angle was related to the orientation of side-chains and carried information about the packing of these groups. All the reference values θ^{ref} were set to zero since our intention was to explore as much conformational space as possible. The fourth collective variable was defined as the 'Debye-Hückel interaction energy'. It calculated the electrostatic interaction between two groups A and B of atoms using an approximation of the Debye-Hückel equation:

$$E_{\text{DH}} = \frac{1}{4\pi\varepsilon_r\varepsilon_0}\sum_{i\in A}\sum_{j\in B}\frac{q_i q_j}{|\mathbf{r}_{ij}|}e^{-\kappa|\mathbf{r}_{ij}|}.$$

The use of this CV was justified by the fact that the studied protein was a dimer and the most stable structures were those that minimize the electrostatic interaction between the two subunits. Group A was defined as formed by atoms belonging to one monomer and group B as formed by atoms of the other monomer.

The height (w), width (σ), and deposition pace (τ) of metadynamics' Gaussian functions were set to the following values:

CV_1 (ϕ):w = 0.05 kJ/mol, σ = 0.50, τ = 20 ps.

CV_2 (ψ):w = 0.05 kJ/mol, σ = 0.50, τ = 20 ps. CV_3 (χ_1):w = 0.05 kJ/mol, σ = 0.30, τ = 20 ps. CV_4 (E_{DH}):w = 0.05 kJ/mol, σ = 0.25, τ = 20 ps.

References

1. Calçada, E. O., Felli, I. C., Hošek, T. & Pierattelli, R. The Heterogeneous Structural Behavior of E7 from HPV16 Revealed by NMR Spectroscopy. *ChemBioChem* **14,** 1876–1882 (2013).

2. Felli, I. C. & Pierattelli, R. Novel methods based on (13)C detection to study intrinsically disordered proteins. *J. Magn. Reson. San Diego Calif 1997* **241,** 115–125 (2014).

3. Brutscher, B. *et al*. NMR Methods for the Study of Instrinsically Disordered Proteins Structure, Dynamics, and Interactions: General Overview and Practical Guidelines. *Adv. Exp. Med. Biol.* **870,** 49–122 (2015).

4. Ottiger, M. & Bax, A. Characterization of magnetically oriented phospholipid micelles for measurement of dipolar couplings in macromolecules. *J. Biomol. NMR* **12,** 361–372 (1998).

5. Rückert, M. & Otting, G. Alignment of Biological Macromolecules in Novel Nonionic Liquid Crystalline Media for NMR Experiments. *J. Am. Chem. Soc.* **122,** 7793–7797 (2000).

6. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314,** 141–151 (1999).

7. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 12562–12566 (2002).

8. Hess, B., Kutzner, C., van der Spoel, D. & Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **4,** 435–447 (2008).

9. Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C. & Bussi, G. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **185,** 604–613 (2014).

10. Cavalli, A., Vendruscolo, M. & Paci, E. Comparison of sequence-based and structure-based energy functions for the reversible folding of a peptide. *Biophys. J.* **88,** 3158–3166 (2005).

11. Fu, B. *et al*. ALMOST: an all atom molecular simulation toolkit for protein structure determination. *J. Comput. Chem.* **35,** 1101–1105 (2014).

12. Wang, R. Y.-R. *et al*. Modeling disordered regions in proteins using Rosetta. *PloS One* **6,** e22060 (2011).

13. Best, R. B. & Hummer, G. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J. Phys. Chem. B* **113,** 9004–9015 (2009).

14. Rauscher, S. *et al*. Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J. Chem. Theory Comput*. **11,** 5513–5524 (2015).

15. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys*. **79,** 926 (1983).

Chapter 4

Conclusions

HPV-16 E7 and HAdV-E1A proteins have been studied for years and despite documented attempts the characterization of these proteins either through NMR or X-ray crystallography failed due to their heterogeneous structural and dynamic properties[54,56,77,287].

Thanks to the recently developed NMR approach to characterize highly heterogeneous proteins[289,290], in combination with clever sample preparation strategies, it was possible to complete the high resolution characterization through NMR of both, HPV16-E7 as well as HAdV-E1A. The information obtained, deposited in the BMRB and in the pEDB, can now be used by the scientific community to answer the many open questions on the molecular determinants of the function of these two distinct proteins that are able to hijack cell regulation by a closely related mechanism. The example of the high resolution study of one of the key post-translational modifications linked to the oncogenic process reported here just represents the first example of many studies that now can be performed taking advantage of the available NMR chemical shift assignment. This information can in turn be used to design innovative drugs that target, instead of well-structured proteins, these two IDPs.

Report of activities

During the three years of doctorate I attended diverse **seminars, conferences and training courses**. With all these activities I earned a total of 45 credits, respecting the minimum credits to be earned (40 credits). The credits obtained distributed by the type of activities is indicated below:

- 32 Internal seminars (3,9 credits)
- 7 Advanced Seminars (5,25 credits)
- 9 PhD program courses (10 credits)
- 4 Courses offered by the Master on Biological Macromolecules (24 credits)
- 2 Oral presentations: "A step further in the NMR characterization of the heterogeneous structural and dynamic behaviour of E7 from HPV16: the C-terminal module" and "The NMR characterization of E7 from HPV16 and its post-translational modifications studies" (2 credits)

Indeed to these activities I had the opportunity to attend **Conferences and training activities** as listed below:

Conferences:

- EMBO workshop on Magnetic Resonance for structural biology, Principina Terra (Grosseto), Italy, June 1-6, 2014;
- IDPbyNMR Final Meeting "High resolution tools to understand the functional role of protein intrinsic disorder", Castiglione della Pescaia, Grosseto, Italy, September 21-26, 2014.

Advanced training course:

- EMBO practical course "Solution and Solid-state NMR of paramagnetic molecules", CERM, University of Florence – Italy, July 13-19, 2014;

Training courses:

- Insurance and risk management according to the University of Florence rules (*Sicurezza Accordo Stato Regione*). Department of Pharmaceutical sciences, University of Florence- Italy, July 11 2013;

- Principle and rules for safety and accident prevention in CERMs laboratories. CERM, University of Florence – Italy, April 20, 2015.

Publications:

During these three years I actively participated in all parts of my PhD project, working on interdisciplinary fields. The ensemble of this work generated three articles:

1. **The highly flexible and heterogeneous nature of HAdV E1A characterized at atomic resolution through NMR** (published in 2016_ DOI 10.1002/chem.201602510 by Chemistry)

   Tomáš Hošek [a], Eduardo O. Calçada [a], Marcela Oliveira Nogueira [a], Michele Salvi [a], Talita D. Pagani [a], Isabella C. Felli [a], Roberta Pierattelli [a]

   [a] Magnetic Resonance Center (CERM) and Department of Chemistry "Ugo Schiff", University of Florence, Italy

2. **Towards the real-time monitoring of HPV-16 E7 phosphorylation events** (submitted to Virology)

   Marcela O. Nogueira [a], Tomáš Hošek [a], Eduardo O. Calcada [a], Francesca Castiglia [a], Paola Massimi [b], Lawrence Banks [b], Isabella C. Felli [a], Roberta Pierattelli [a]

   [a] Magnetic Resonance Center (CERM) and Department of Chemistry "Ugo Schiff", University of Florence, Italy

   [b] International Centre for Genetic Engineering and Biotechnology (ICGEB),, Trieste, Italy

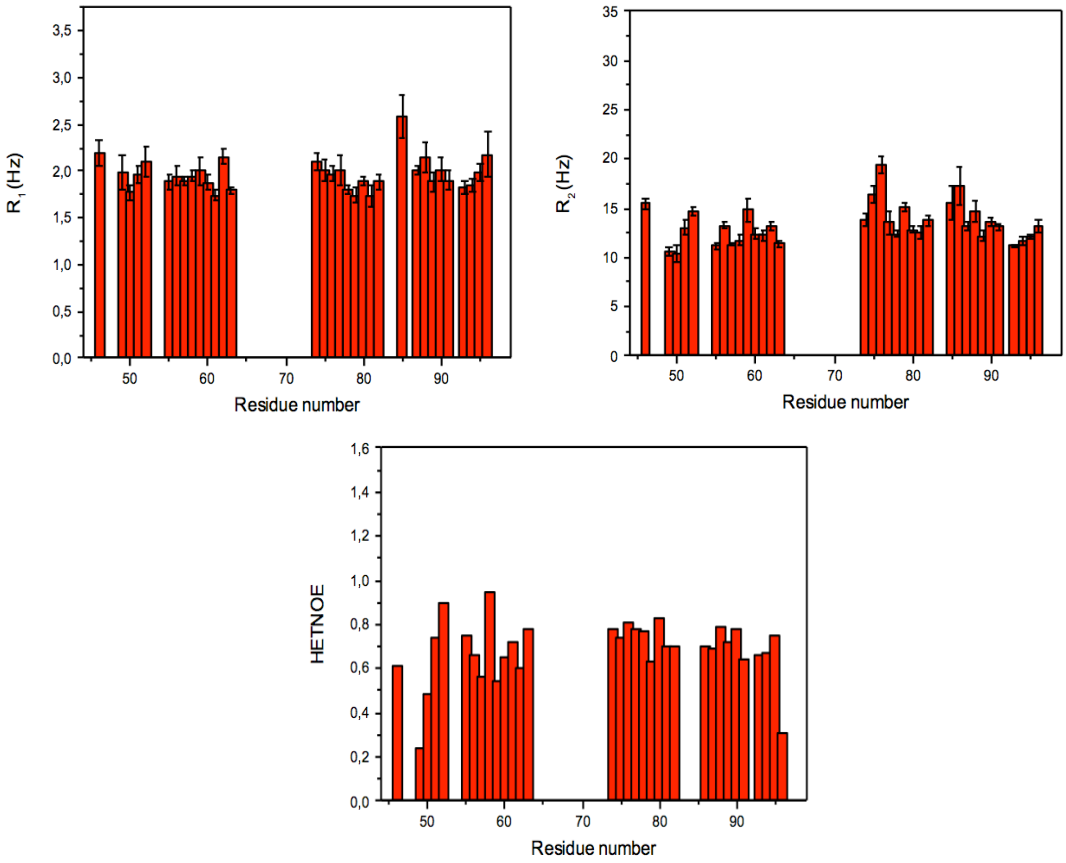
3. **The structure and dynamics of HPV-16 E7 described by NMR-guided metadynamics** (in preparation)

   Magnetic Resonance Center (CERM) and Department of Chemistry "Ugo Schiff", University of Florence, Italy

Department of Chemistry, University of Cambridge, Cambridge, United Kingdom

Appendices

Relaxation rates of E7CR3



Relaxation rates of ^{15}N labelled E7CR3 measured at 11.8 T on a Bruker instrument. Sample conditions: 250 μM protein in 50 mM HEPES, pH 8.5; 50 mM KCl, 5 mM DTT.

Sequence alignment of E1A289 from HAdV-2 and HAdV-5

```
HAdV-2     MRHIICHGGVITEEMAASLLDQLIEEVLADNLPPPSHFEPPTLHELYDLDVTAPEDPNEE
HAdV-2/5   MRHIICHGGVITEEMAASLLDQLIEEVLADNLPPPSHFEPPTLHELYDLDVTAPEDPNEE
HAdV-5     MRHIICHGGVITEEMAASLLDQLIEEVLADNLPPPSHFEPPTLHELYDLDVTAPEDPNEE
           ************************************************************

HAdV-2     AVSQIFPESVMLAVQEGIDLFTFPPAPGSPEPPHLSRQPEQPEQRALGPVSMPNLVPEVI
HAdV-2/5   AVSQIFPESVMLAVQEGIDLFTFPPAPGSPEPPHLSRQPEQPEQRALGPVSMPNLVPEVI
HAdV-5     AVSQIFPDSVMLAVQEGIDLLTFPPAPGSPEPPHLSRQPEQPEQRALGPVSMPNLVPEVI
           *******:************:*****************************************

HAdV-2     DLTCHEAGFPPSDDEDEEGEEFVLDYVEHPGHGCRSCHYHRRNTGDPDIMCSLCYMRTCG
HAdV-2/5   DLTCHEAGFPPSDDEDEEGEEFVLDYVGHPGHGCRSCHYHRRNTGDPDIMCSLCYMRTCG
HAdV-5     DLTCHEAGFPPSDDEDEEGEEFVLDYVEHPGHGCRSCHYHRRNTGDPDIMCSLCYMRTCG
           ***************************:**********************************

HAdV-2     MFVYSPVSEPEPEPEPEPEPARPTRRPKLVPAILRRPTSPVSRECNSSTDSCDSGPSNTP
HAdV-2/5   MFVYSPVSEPEPEPEPEPEPARPTRRPKLVPAILRRPTSPVSRECNSSTDSCDSGPSNTP
HAdV-5     MFVYSPVSEPEPEPEPEPEPARPTRRPKMAPAILRRPTSPVSRECNSSTDSCDSGPSNTP
           ****************************::*******************************

HAdV-2     PEIHPVVPLCPIKPVAVRVGGRRQAVECIEDLLNESGQPLDLSCKRPRP
HAdV-2/5   PEIHPVVPLCPIKPVAVRVGGRRQAVECIEDLLNEPGQPLDLSCKRPRP
HAdV-5     PEIHPVVPLCPIKPVAVRVGGRRQAVECIEDLLNEPGQPLDLSCKRPRP
           ***********************************:**************
```

The sequence alignment of E1A289 from the HAdV-2, the HAdV-5 and the hybrid HAdV-2/5 (studied in the present work). Asterisks (*) indicate the same residue while colons (:) indicate the presence of a different residue among the three sequences. E1A289 from the hybrid HAdV-2/5 is essentially the same as E1A289 from HAdV-2, with the exception of the residues highlighted in Red (Gly148 and Pro276).

Sequence alignment of E1A243 and E1A289

```
E1A243    MRHIICHGGVITEEMAASLLDQLIEEVLADNLPPPSHFEPPTLHELYDLDVTAPEDPNEE
E1A289    MRHIICHGGVITEEMAASLLDQLIEEVLADNLPPPSHFEPPTLHELYDLDVTAPEDPNEE
          ************************************************************
E1A243    AVSQIFPESVMLAVQEGIDLFTFPPAPGSPEPPHLSRQPEQPEQRALGPVSMPNLVPEVI
E1A289    AVSQIFPESVMLAVQEGIDLFTFPPAPGSPEPPHLSRQPEQPEQRALGPVSMPNLVPEVI
          ************************************************************
E1A243    DLTCHEAGFPPSDDEDEEG-----------------------------------------
E1A289    DLTCHEAGFPPSDDEDEEGEEFVLDYVGHPGHGCRSCHYHRRNTGDPDIMCSLCYMRTCG
          ******************
E1A243    -----PVSEPEPEPEPEPEPARPTRRPKLVPAILRRPTSPVSRECNSSTDSCDSGPSNTP
E1A289    MFVYSPVSEPEPEPEPEPEPARPTRRPKLVPAILRRPTSPVSRECNSSTDSCDSGPSNTP
               ********************************************************
E1A243    PEIHPVVPLCPIKPVAVRVGGRRQAVECIEDLLNEPGQPLDLSCKRPRP
E1A289    PEIHPVVPLCPIKPVAVRVGGRRQAVECIEDLLNEPGQPLDLSCKRPRP
          *************************************************
```

The sequence alignment of E1A243 and E1A289 from the hybrid HAdV-2/5 studied in this work. Both sequences are the same except by the presence of the CR3 domain (46 residues), present in E1A289 but not in E1A243. Asterisks (*) indicate the same residues while hyphens (-) indicate missing residues in one of the two sequences.

Table AP1- Amino acid content of HPV-16 E7, E7CR3 and E1ACR3

Amino acid	Number of residues and their percentage respect to the whole protein		
	HPV-16 E7	**E7CR3**	**E1ACR3**
Ala (A)	3 (3.1%)	2 (2.7%)	0 0.0%
Arg (R)	3 (3.1%)	3 (5.6%)	4 7.3%
Asn (N)	2 (2.0%)	1 (1.9%)	1 1.8%
Asp (D)	10 (10.2%)	4 (7.4%)	4 7.3%
Cys (C)	7 (7.1%)	6 11.1%	5 9.1%
Gln (Q)	5 (5.1%)	2 3.7%	0 0.0%
Glu (E)	9 (9.2%)	2 3.7%	6 10.9%
Gly (G)	5 (5.1%)	2 3.7%	5 9.1%
His (H)	4 (4.1%)	2 3.7%	4 7.3%
Ile (I)	5 (5.1%)	4 7.4%	1 1.8%
Leu (L)	11 (11.2%)	6 11.1%	2 3.6%
Lys (K)	2 (2.0%)	**2 3.7%**	0 0.0%
Met (M)	3 (3.1%)	2 3.7%	4 7.3%
Phe (F)	1 (1.0%)	1 1.9%	2 3.6%
Pro (P)	6 (6.1%)	2 3.(7%)	3 5.5%
Ser (S)	5 (5.1%)	3 5.(6%)	4 7.3%
Thr (T)	9 (9.2%)	5 (9.3%)	2 3.6%
Trp (W)	0 (0.0%)	0 (0.0%)	0 0.0%
Tyr (Y)	4 (4.1%)	1 (1.9%)	4 7.3%
Val (V)	4 (4.1%)	4 (7.4%)	4 7.3%

Table AP2 - Protein parameter calculated by ProtParam tool

Parameters	HPV-16 E7	E7CR3	E1ACR3
Number of amino acids	98	54	55
Molecular weight	11022.3	6043.1	6391.1
Theoretical isoelectric point	4.20	5.99	4.85
Extinction coefficients [a]	5960	1490	5960

(a) (M-1 cm-1, at 280 nm) Assuming that all Cys residues are reduced

References

1. Gelderblom, H. R. in *Medical Microbiology* (ed. Baron, S.) (University of Texas Medical Branch at Galveston, 1996).

2. Bouvard, V. *et al*. A review of human carcinogens--Part B: biological agents. *Lancet Oncol*. **10,** 321–322 (2009).

3. de Martel, C. *et al*. Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *Lancet Oncol*. **13,** 607–615 (2012).

4. zur Hausen, H. The search for infectious causes of human cancers: where and why (Nobel lecture). *Angew. Chem. Int. Ed Engl*. **48,** 5798–5808 (2009).

5. Ekman, D., Sara, L., Åsa, K. B. & Arne, E. What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? *BioMed Central* (2006).

6. Tsai, C.-J., Ma, B. & Nussinov, R. Protein-protein interaction networks: how can a hub protein bind so many different partners? *Trends Biochem. Sci*. **34,** 594–600 (2009).

7. Bloomer, A. C., Champness, J. N., Bricogne, G., Staden, R. & Klug, A. Protein disk of tobacco mosaic virus at 2.8 Å resolution showing the interactions within and between subunits. *Nature* **276,** 362–368 (1978).

8. Bode, W., Schwager, P. & Huber, R. The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding. *J. Mol. Biol*. **118,** 99–112 (1978).

9. Lemieux, R. U. & Spohr, U. How Emil Fischer was led to the lock and key concept for enzyme specificity. *Adv. Carbohydr. Chem. Biochem*. **50,** 1–20 (1994).

10. Haynes, C. *et al*. Intrinsic Disorder Is a Common Feature of Hub Proteins from Four Eukaryotic Interactomes. *PLoS Comput. Biol*. **2,** e100 (2006).

11. Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M. & Uversky, V. N. Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J*. **272,** 5129–5148 (2005).

12. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J. Mol. Recognit. JMR* **18,** 343–384 (2005).

13. Dunker, A. K. *et al*. Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac. Symp. Biocomput. Pac. Symp. Biocomput*. 473–484 (1998).

14. Uversky, V. N., Gillespie, J. R. & Fink, A. L. Why are 'natively unfolded' proteins unstructured under physiologic conditions? *Proteins* **41,** 415–427 (2000).

15. Dunker, A. K. *et al*. Intrinsically disordered protein. *J. Mol. Graph. Model*. **19,** 26–59 (2001).

16. Williams, R. M. *et al*. The protein non-folding problem: amino acid determinants of intrinsic order and disorder. *Pac. Symp. Biocomput. Pac. Symp. Biocomput*. 89–100 (2001).

17. Campen, A. *et al*. TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein Pept. Lett*. **15,** 956–963 (2008).

18. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradović, Z. Intrinsic disorder and protein function. *Biochemistry (Mosc.)* **41,** 6573–6582 (2002).

19. Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol*. **293,** 321–331 (1999).

20. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol*. **6,** 197–208 (2005).

21. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradović, Z. & Dunker, A. K. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol*. **323,** 573–584 (2002).

22. Galea, C. A., Wang, Y., Sivakolundu, S. G. & Kriwacki, R. W. Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. *Biochemistry (Mosc.)* **47,** 7598–7609 (2008).

23. Babu, M. M., van der Lee, R., de Groot, N. S. & Gsponer, J. Intrinsically disordered proteins: regulation and disease. *Curr. Opin. Struct. Biol*. **21,** 432–440 (2011).

24. Guharoy, M., Szabo, B., Contreras Martos, S., Kosol, S. & Tompa, P. Intrinsic structural disorder in cytoskeletal proteins. *Cytoskelet. Hoboken NJ* **70,** 550–571 (2013).

25. Uversky, V. N. The alphabet of intrinsic disorder: II. Various roles of glutamic acid in ordered and intrinsically disordered proteins. *Intrinsically Disord. Proteins* **1,** e24684 (2013).

26. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *J. Mol. Biol.* **337,** 635–645 (2004).

27. Xue, B., Dunker, A. K. & Uversky, V. N. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J. Biomol. Struct. Dyn.* **30,** 137–149 (2012).

28. Pelka, P., Ablack, J. N. G., Fonseca, G. J., Yousef, A. F. & Mymryk, J. S. Intrinsic structural disorder in adenovirus E1A: a viral molecular hub linking multiple diverse processes. *J. Virol.* **82,** 7252–7263 (2008).

29. Uversky, V. N. & Dunker, A. K. Understanding protein non-folding. *Biochim. Biophys. Acta BBA - Proteins Proteomics* **1804,** 1231–1264 (2010).

30. Lian, L. Y. NMR structural studies of glutathione S-transferase. *Cell. Mol. Life Sci. CMLS* **54,** 359–362 (1998).

31. Sidote, D. J. & Hoffman, D. W. NMR structure of an archaeal homologue of ribonuclease P protein Rpp29. *Biochemistry (Mosc.)* **42,** 13541–13550 (2003).

32. Mizutani, H. *et al.* Systematic study on crystal-contact engineering of diphthine synthase: influence of mutations at crystal-packing regions on X-ray diffraction quality. *Acta Crystallogr. D Biol. Crystallogr.* **64,** 1020–1033 (2008).

33. Bosch, F. X. *et al.* Prevalence of human papillomavirus in cervical cancer: a worldwide perspective. International biological study on cervical cancer (IBSCC) Study Group. *J. Natl. Cancer Inst.* **87,** 796–802 (1995).

34. Bernard, H.-U. *et al.* Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology* **401,** 70–79 (2010).

35. zur Hausen, H. Papillomaviruses and cancer: from basic studies to clinical application. *Nat. Rev. Cancer* **2,** 342–350 (2002).

36. Muñoz, N. *et al.* Epidemiologic Classification of Human Papillomavirus Types Associated with Cervical Cancer. *N. Engl. J. Med.* **348,** 518–527 (2003).

37. Arbyn, M. *et al*. Worldwide burden of cervical cancer in 2008. *Ann. Oncol.* **22,** 2675–2686 (2011).

38. Uversky, V. N., Roman, A., Oldfield, C. J. & Dunker, A. K. Protein intrinsic disorder and human papillomaviruses: increased amount of disorder in E6 and E7 oncoproteins from high risk HPVs. *J. Proteome Res.* **5,** 1829–1842 (2006).

39. Doorbar, J. Molecular biology of human papillomavirus infection and cervical cancer. *Clin. Sci.* **110,** 525–541 (2006).

40. Modis, Y. Atomic model of the papillomavirus capsid. *EMBO J.* **21,** 4754–4762 (2002).

41. Hughes, F. J. & Romanos, M. A. E1 protein of human papillomavirus is a DNA helicase/ATPase. *Nucleic Acids Res.* **21,** 5817–5823 (1993).

42. Ustav, M. & Stenlund, A. Transient replication of BPV-1 requires two viral polypeptides encoded by the E1 and E2 open reading frames. *EMBO J.* **10,** 449–457 (1991).

43. Wilson, R., Fehrmann, F. & Laimins, L. A. Role of the E1--E4 protein in the differentiation-dependent life cycle of human papillomavirus type 31. *J. Virol.* **79,** 6732–6740 (2005).

44. Brown, D. R. *et al*. The human papillomavirus type 11 E1--E4 protein is a transglutaminase 3 substrate and induces abnormalities of the cornified cell envelope. *Virology* **345,** 290–298 (2006).

45. Davy, C. E. *et al*. HPV16 E1--E4 protein is phosphorylated by Cdk2/cyclin A and relocalizes this complex to the cytoplasm. *Virology* **349,** 230–244 (2006).

46. Straight, S. W., Hinkle, P. M., Jewers, R. J. & McCance, D. J. The E5 oncoprotein of human papillomavirus type 16 transforms fibroblasts and effects the downregulation of the epidermal growth factor receptor in keratinocytes. *J. Virol.* **67,** 4521–4532 (1993).

47. Fehrmann, F., Klumpp, D. J. & Laimins, L. A. Human papillomavirus type 31 E5 protein supports cell cycle progression and activates late viral functions upon epithelial differentiation. *J. Virol.* **77,** 2819–2831 (2003).

48. Chemes, L. B., Sánchez, I. E., Smal, C. & de Prat-Gay, G. Targeting mechanism of the retinoblastoma tumor suppressor by a prototypical viral

oncoprotein. Structural modularity, intrinsic disorder and phosphorylation of human papillomavirus E7. *FEBS J*. **277,** 973–988 (2010).

49. Münger, K., Scheffner, M., Huibregtse, J. M. & Howley, P. M. Interactions of HPV E6 and E7 oncoproteins with tumour suppressor gene products. *Cancer Surv*. **12,** 197–217 (1992).

50. Halbert, C. L., Demers, G. W. & Galloway, D. A. The E7 gene of human papillomavirus type 16 is sufficient for immortalization of human epithelial cells. *J. Virol*. **65,** 473–478 (1991).

51. Thomas, M. *et al*. Human papillomaviruses, cervical cancer and cell polarity. *Oncogene* **27,** 7018–7030 (2008).

52. Phelps, W. C., Münger, K., Yee, C. L., Barnes, J. A. & Howley, P. M. Structure-function analysis of the human papillomavirus type 16 E7 oncoprotein. *J. Virol*. **66,** 2418–2427 (1992).

53. Roman, A. & Munger, K. The papillomavirus E7 proteins. *Virology* **445,** 138–168 (2013).

54. Liu, X., Clements, A., Zhao, K. & Marmorstein, R. Structure of the human Papillomavirus E7 oncoprotein and its mechanism for inactivation of the retinoblastoma tumor suppressor. *J. Biol. Chem*. **281,** 578–586 (2006).

55. García-Alai, M. M., Alonso, L. G. & de Prat-Gay, G. The N-Terminal Module of HPV16 E7 Is an Intrinsically Disordered Domain That Confers Conformational and Recognition Plasticity to the Oncoprotein. *Biochemistry (Mosc.)* **46,** 10405–10412 (2007).

56. Calçada, E. O., Felli, I. C., Hošek, T. & Pierattelli, R. The Heterogeneous Structural Behavior of E7 from HPV16 Revealed by NMR Spectroscopy. *ChemBioChem* **14,** 1876–1882 (2013).

57. Lee, J. O., Russo, A. A. & Pavletich, N. P. Structure of the retinoblastoma tumour-suppressor pocket domain bound to a peptide from HPV E7. *Nature* **391,** 859–865 (1998).

58. Münger, K. *et al*. Complex formation of human papillomavirus E7 proteins with the retinoblastoma tumor suppressor gene product. *EMBO J*. **8,** 4099–4105 (1989).

59. Hatakeyama, M. & Weinberg, R. A. The role of RB in cell cycle control. *Prog. Cell Cycle Res.* **1,** 9–19 (1995).

60. Dyson, N., Howley, P. M., Münger, K. & Harlow, E. The human papilloma virus-16 E7 oncoprotein is able to bind to the retinoblastoma gene product. *Science* **243,** 934–937 (1989).

61. Davies, R., Hicks, R., Crook, T., Morris, J. & Vousden, K. Human papillomavirus type 16 E7 associates with a histone H1 kinase and with p107 through sequences necessary for transformation. *J. Virol.* **67,** 2521–2528 (1993).

62. Tommasino, M. *et al.* HPV16 E7 protein associates with the protein kinase p33CDK2 and cyclin A. *Oncogene* **8,** 195–202 (1993).

63. Zerfass-Thome, K. *et al.* Inactivation of the cdk inhibitor p27KIP1 by the human papillomavirus type 16 E7 oncoprotein. *Oncogene* **13,** 2323–2330 (1996).

64. Funk, J. O. *et al.* Inhibition of CDK activity and PCNA-dependent DNA replication by p21 is blocked by interaction with the HPV-16 E7 oncoprotein. *Genes Dev.* **11,** 2090–2100 (1997).

65. Massimi, P., Pim, D. & Banks, L. Human papillomavirus type 16 E7 binds to the conserved carboxy-terminal region of the TATA box binding protein and this contributes to E7 transforming activity. *J. Gen. Virol.* **78 (Pt 10),** 2607–2613 (1997).

66. Bernat, A., Avvakumov, N., Mymryk, J. S. & Banks, L. Interaction between the HPV E7 oncoprotein and the transcriptional coactivator p300. *Oncogene* **22,** 7871–7881 (2003).

67. Hwang, S. G., Lee, D., Kim, J., Seo, T. & Choe, J. Human papillomavirus type 16 E7 binds to E2F1 and activates E2F1-driven transcription in a retinoblastoma protein-independent manner. *J. Biol. Chem.* **277,** 2923–2930 (2002).

68. Genovese, N. J., Banerjee, N. S., Broker, T. R. & Chow, L. T. Casein Kinase II Motif-Dependent Phosphorylation of Human Papillomavirus E7 Protein Promotes p130 Degradation and S-Phase Induction in Differentiated Human Keratinocytes. *J. Virol.* **82,** 4862–4873 (2008).

69. Massimi, P. & Banks, L. Differential Phosphorylation of the HPV-16 E7 Oncoprotein during the Cell Cycle. *Virology* **276,** 388–394 (2000).

70. Jansma, A. L. *et al*. The High-Risk HPV16 E7 Oncoprotein Mediates Interaction between the Transcriptional Coactivator CBP and the Retinoblastoma Protein pRb. *J. Mol. Biol*. **426,** 4030–4048 (2014).

71. Chien, W. M., Parker, J. N., Schmidt-Grimminger, D. C., Broker, T. R. & Chow, L. T. Casein kinase II phosphorylation of the human papillomavirus-18 E7 protein is critical for promoting S-phase entry. *Cell Growth Differ. Mol. Biol. J. Am. Assoc. Cancer Res*. **11,** 425–435 (2000).

72. Todorovic, B. *et al*. Conserved Region 3 of Human Papillomavirus 16 E7 Contributes to Deregulation of the Retinoblastoma Tumor Suppressor. *J. Virol*. **86,** 13313–13323 (2012).

73. Barbosa, M. S., Lowy, D. R. & Schiller, J. T. Papillomavirus polypeptides E6 and E7 are zinc-binding proteins. *J. Virol*. **63,** 1404–1407 (1989).

74. McIntyre, M. C., Frattini, M. G., Grossman, S. R. & Laimins, L. A. Human papillomavirus type 18 E7 protein requires intact Cys-X-X-Cys motifs for zinc binding, dimerization, and transformation but not for Rb binding. *J. Virol*. **67,** 3142–3150 (1993).

75. Clements, A., Johnston, K., Mazzarelli, J. M., Ricciardi, R. P. & Marmorstein, R. Oligomerization Properties of the Viral Oncoproteins Adenovirus E1A and Human Papillomavirus E7 and Their Complexes with the Retinoblastoma Protein †. *Biochemistry (Mosc.)* **39,** 16033–16045 (2000).

76. Clemens, K. E., Brent, R., Gyuris, J. & MüNger, K. Dimerization of the Human Papillomavirus E7 Oncoproteinin Vivo. *Virology* **214,** 289–293 (1995).

77. Ohlenschläger, O. *et al*. Solution structure of the partially folded high-risk human papilloma virus 45 oncoprotein E7. *Oncogene* **25,** 5953–5959 (2006).

78. Todorovic, B. *et al*. Systematic Analysis of the Amino Acid Residues of Human Papillomavirus Type 16 E7 Conserved Region 3 Involved in Dimerization and Transformation. *J. Virol*. **85,** 10048–10057 (2011).

79. Noval, M. G. *et al*. Conformational Dissection of a Viral Intrinsically Disordered Domain Involved in Cellular Transformation. *PLoS ONE* **8,** e72760 (2013).

80. Stewart, P. L., Fuller, S. D. & Burnett, R. M. Difference imaging of adenovirus: bridging the resolution gap between X-ray crystallography and electron microscopy. *EMBO J.* **12,** 2589–2599 (1993).

81. Davison, A. J., Benko, M. & Harrach, B. Genetic content and evolution of adenoviruses. *J. Gen. Virol.* **84,** 2895–2908 (2003).

82. Vellinga, J. The adenovirus capsid: major progress in minor proteins. *J. Gen. Virol.* **86,** 1581–1588 (2005).

83. Rowe, W. P., Huebner, R. J., Gilmore, L. K., Parrott, R. H. & Ward, T. G. Isolation of a Cytopathogenic Agent from Human Adenoids Undergoing Spontaneous Degeneration in Tissue Culture. *Exp. Biol. Med.* **84,** 570–573 (1953).

84. Benkö, M. & Harrach, B. Molecular evolution of adenoviruses. *Curr. Top. Microbiol. Immunol.* **272,** 3–35 (2003).

85. Jones, M. S. *et al.* New adenovirus species found in a patient presenting with gastroenteritis. *J. Virol.* **81,** 5978–5984 (2007).

86. Ishiko, H. *et al.* Novel human adenovirus causing nosocomial epidemic keratoconjunctivitis. *J. Clin. Microbiol.* **46,** 2002–2008 (2008).

87. Zhang, S.-Y. *et al.* Fatal pneumonia cases caused by human adenovirus 55 in immunocompetent adults. *Infect. Dis. Lond. Engl.* **48,** 40–47 (2016).

88. Trentin, J. J., Van Hoosier, G. L. & Samper, L. The oncogenicity of human adenoviruses in hamsters. *Proc. Soc. Exp. Biol. Med. Soc. Exp. Biol. Med. N. Y. N* **127,** 683–689 (1968).

89. Trentin, J. J., Yabe, Y. & Taylor, G. The quest for human cancer viruses. *Science* **137,** 835–841 (1962).

90. Freeman, A. E. *et al.* Transformation of primary rat embryo cells by adenovirus type 2. *Proc. Natl. Acad. Sci. U. S. A.* **58,** 1205–1212 (1967).

91. Ilan, J. *Translational regulation of gene expression.* (Plenum Press, 1987).

92. Farley, D. C., Brown, J. L. & Leppard, K. N. Activation of the early-late switch in adenovirus type 5 major late transcription unit expression by L4 gene products. *J. Virol.* **78,** 1782–1791 (2004).

93. Houweling, A., van den Elsen, P. J. & van der Eb, A. J. Partial transformation of primary rat cells by the leftmost 4.5% fragment of adenovirus 5 DNA. *Virology* **105,** 537–550 (1980).

94. Flint, J. & Shenk, T. Adenovirus E1A protein paradigm viral transactivator. *Annu. Rev. Genet.* **23,** 141–161 (1989).

95. Frisch, S. M. Antioncogenic effect of adenovirus E1A in human tumor cells. *Proc. Natl. Acad. Sci. U. S. A.* **88,** 9077–9081 (1991).

96. Radko, S., Jung, R., Olanubi, O. & Pelka, P. Effects of Adenovirus Type 5 E1A Isoforms on Viral Replication in Arrested Human Cells. *PloS One* **10,** e0140124 (2015).

97. Kimelman, D., Miller, J. S., Porter, D. & Roberts, B. E. E1a regions of the human adenoviruses and of the highly oncogenic simian adenovirus 7 are closely related. *J. Virol.* **53,** 399–409 (1985).

98. Avvakumov, N., Kajon, A. E., Hoeben, R. C. & Mymryk, J. S. Comprehensive sequence analysis of the E1A proteins of human and simian adenoviruses. *Virology* **329,** 477–492 (2004).

99. Chinnadurai, G. Opposing oncogenic activities of small DNA tumor virus transforming proteins. *Trends Microbiol.* **19,** 174–183 (2011).

100. Romero, P., Obradovic, Z. & Dunker, A. K. Natively disordered proteins: functions and predictions. *Appl. Bioinformatics* **3,** 105–113 (2004).

101. Xue, B., Dunbrack, R. L., Williams, R. W., Dunker, A. K. & Uversky, V. N. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta* **1804,** 996–1010 (2010).

102. Davey, N. E., Travé, G. & Gibson, T. J. How viruses hijack cell regulation. *Trends Biochem. Sci.* **36,** 159–169 (2011).

103. Goodman, R. H. & Smolik, S. CBP/p300 in cell growth, transformation, and development. *Genes Dev.* **14,** 1553–1577 (2000).

104. Barbeau, D., Charbonneau, R., Whalen, S. G., Bayley, S. T. & Branton, P. E. Functional interactions within adenovirus E1A protein complexes. *Oncogene* **9,** 359–373 (1994).

105. Deleu, L., Shellard, S., Alevizopoulos, K., Amati, B. & Land, H. Recruitment of TRRAP required for oncogenic transformation by E1A. *Oncogene* **20,** 8270–8275 (2001).

106. McMahon, S. B., Van Buskirk, H. A., Dugan, K. A., Copeland, T. D. & Cole, M. D. The novel ATM-related protein TRRAP is an essential cofactor for the c-Myc and E2F oncoproteins. *Cell* **94,** 363–374 (1998).

107. Whyte, P. *et al*. Association between an oncogene and an anti-oncogene: the adenovirus E1A proteins bind to the retinoblastoma gene product. *Nature* **334,** 124–129 (1988).

108. Ikeda, M. A. & Nevins, J. R. Identification of distinct roles for separate E1A domains in disruption of E2F complexes. *Mol. Cell. Biol.* **13,** 7029–7035 (1993).

109. Dyson, N., Guida, P., McCall, C. & Harlow, E. Adenovirus E1A makes two distinct contacts with the retinoblastoma protein. *J. Virol.* **66,** 4606–4611 (1992).

110. Ansieau, S. & Leutz, A. The conserved Mynd domain of BS69 binds cellular and oncoviral proteins through a common PXLXP motif. *J. Biol. Chem.* **277,** 4906–4910 (2002).

111. Hateboer, G. *et al*. BS69, a novel adenovirus E1A-associated protein that inhibits E1A transactivation. *EMBO J.* **14,** 3159–3169 (1995).

112. Masselink, H. & Bernards, R. The adenovirus E1A binding protein BS69 is a corepressor of transcription through recruitment of N-CoR. *Oncogene* **19,** 1538–1546 (2000).

113. Johnson, E. S. Protein modification by SUMO. *Annu. Rev. Biochem.* **73,** 355–382 (2004).

114. Gareau, J. R. & Lima, C. D. The SUMO pathway: emerging mechanisms that shape specificity, conjugation and recognition. *Nat. Rev. Mol. Cell Biol.* **11,** 861–871 (2010).

115. Zhang, X. *et al*. The targeting of the proteasomal regulatory subunit S2 by adenovirus E1A causes inhibition of proteasomal activity and increased p53 expression. *J. Biol. Chem.* **279,** 25122–25133 (2004).

116. Love, I. M., Shi, D. & Grossman, S. R. p53 Ubiquitination and proteasomal degradation. *Methods Mol. Biol. Clifton NJ* **962,** 63–73 (2013).

117. Culp, J. S. *et al*. The 289-amino acid E1A protein of adenovirus binds zinc in a region that is important for trans-activation. *Proc. Natl. Acad. Sci. U. S. A.* **85,** 6450–6454 (1988).

118. Lillie, J. W. & Green, M. R. Transcription activation by the adenovirus E1a protein. *Nature* **338,** 39–44 (1989).

119. Bayley, S. T. & Mymryk, J. S. Adenovirus e1a proteins and transformation (review). *Int. J. Oncol.* **5,** 425–444 (1994).

120. Flint, J. & Shenk, T. Viral transactivating proteins. *Annu. Rev. Genet.* **31,** 177–212 (1997).

121. Geisberg, J. V., Lee, W. S., Berk, A. J. & Ricciardi, R. P. The zinc finger region of the adenovirus E1A transactivating domain complexes with the TATA box binding protein. *Proc. Natl. Acad. Sci. U. S. A.* **91,** 2488–2492 (1994).

122. Boyer, T. G., Martin, M. E., Lees, E., Ricciardi, R. P. & Berk, A. J. Mammalian Srb/Mediator complex is targeted by adenovirus E1A protein. *Nature* **399,** 276–279 (1999).

123. Jelsma, T. N. *et al*. Use of deletion and point mutants spanning the coding region of the adenovirus 5 E1A gene to define a domain that is essential for transcriptional activation. *Virology* **163,** 494–502 (1988).

124. Fahnestock, M. L. & Lewis, J. B. Genetic dissection of the transactivating domain of the E1a 289R protein of adenovirus type 2. *J. Virol.* **63,** 1495–1504 (1989).

125. Webster, L. C. & Ricciardi, R. P. trans-dominant mutants of E1A provide genetic evidence that the zinc finger of the trans-activating domain binds a transcription factor. *Mol. Cell. Biol.* **11,** 4287–4296 (1991).

126. Webster, L. C. *et al*. Conversion of the E1A Cys4 zinc finger to a nonfunctional His2,Cys2 zinc finger by a single point mutation. *Proc. Natl. Acad. Sci. U. S. A.* **88,** 9989–9993 (1991).

127. Ciechanover, A. The ubiquitin-proteasome pathway: on protein death and cell life. *EMBO J.* **17,** 7151–7160 (1998).

128. Subramanian, T., Malstrom, S. E. & Chinnadurai, G. Requirement of the C-terminal region of adenovirus E1a for cell transformation in cooperation with E1b. *Oncogene* **6,** 1171–1173 (1991).

129. Subramanian, T., La Regina, M. & Chinnadurai, G. Enhanced ras oncogene mediated cell transformation and tumorigenesis by adenovirus 2 mutants lacking the C-terminal region of E1a protein. *Oncogene* **4,** 415–420 (1989).

130. Boyd, J. M. *et al*. A region in the C-terminus of adenovirus 2/5 E1a protein is required for association with a cellular phosphoprotein and important for the negative modulation of T24-ras mediated transformation, tumorigenesis and metastasis. *EMBO J*. **12,** 469–478 (1993).

131. Schaeper, U. *et al*. Molecular cloning and characterization of a cellular phosphoprotein that interacts with a conserved C-terminal domain of adenovirus E1A involved in negative modulation of oncogenic transformation. *Proc. Natl. Acad. Sci. U. S. A.* **92,** 10467–10471 (1995).

132. Chinnadurai, G. CtBP, an Unconventional Transcriptional Corepressor in Development and Oncogenesis. *Mol. Cell* **9,** 213–224 (2002).

133. Chinnadurai, G. in *GtBP Family Proteins* 1–17 (Springer New York, 2007).

134. Zhao, L.-J., Subramanian, T., Vijayalingam, S. & Chinnadurai, G. PLDLS-dependent interaction of E1A with CtBP: regulation of CtBP nuclear localization and transcriptional functions. *Oncogene* **26,** 7544–7551 (2007).

135. Human Genome Sequencing Consortium, I. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).

136. Pertea, M. & Salzberg, S. L. Between a chicken and a grape: estimating the number of human genes. *Genome Biol*. **11,** 206 (2010).

137. Jensen, O. N. Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Curr. Opin. Chem. Biol*. **8,** 33–41 (2004).

138. Ayoubi, T. A. & Van De Ven, W. J. Regulation of gene expression by alternative promoters. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol*. **10,** 453–460 (1996).

139. Black, D. L. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell* **103,** 367–370 (2000).

140. Beltrao, P., Bork, P., Krogan, N. J. & van Noort, V. Evolution and functional cross-talk of protein post-translational modifications. *Mol. Syst. Biol.* **9,** n/a-n/a (2013).

141. Seet, B. T., Dikic, I., Zhou, M.-M. & Pawson, T. Reading protein modifications with interaction domains. *Nat. Rev. Mol. Cell Biol.* **7,** 473–483 (2006).

142. Iakoucheva, L. M. *et al*. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* **32,** 1037–1049 (2004).

143. Radivojac, P. *et al*. Intrinsic Disorder and Functional Proteomics. *Biophys. J.* **92,** 1439–1456 (2007).

144. Xie, H. *et al*. Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. *J. Proteome Res.* **6,** 1917–1932 (2007).

145. Gao, J. & Xu, D. Correlation between posttranslational modification and intrinsic disorder in protein. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* 94–103 (2012).

146. Minguez, P. *et al*. Deciphering a global network of functionally associated post-translational modifications. *Mol. Syst. Biol.* **8,** 599 (2012).

147. Dumont, D. J. & Branton, P. E. Phosphorylation of adenovirus E1A proteins by the p34cdc2 protein kinase. *Virology* **189,** 111–120 (1992).

148. Mal, A., Piotrkowski, A. & Harter, M. L. Cyclin-dependent kinases phosphorylate the adenovirus E1A protein, enhancing its ability to bind pRb and disrupt pRb-E2F complexes. *J. Virol.* **70,** 2911–2921 (1996).

149. Tremblay, M. L., McGlade, C. J., Gerber, G. E. & Branton, P. E. Identification of the phosphorylation sites in early region 1A proteins of adenovirus type 5 by amino acid sequencing of peptide fragments. *J. Biol. Chem.* **263,** 6375–6383 (1988).

150. Tremblay, M. L., Dumont, D. J. & Branton, P. E. Analysis of phosphorylation sites in the exon 1 region of E1A proteins of human adenovirus type 5. *Virology* **169,** 397–407 (1989).

151. Dumont, D. J., Marcellus, R. C., Bayley, S. T. & Branton, P. E. Role of phosphorylation near the amino terminus of adenovirus type 5 early region 1A proteins. *J. Gen. Virol.* **74 (Pt 4),** 583–595 (1993).

152. Tsukamoto, A. S., Pnticelli, A., Berk, A. J. & Gaynor, R. B. Genetic mapping of a major site of phosphorylation in adenovirus type 2 E1A proteins. *J. Virol.* **59,** 14–22 (1986).

153. Whalen, S. G., Marcellus, R. C., Barbeau, D. & Branton, P. E. Importance of the Ser-132 Phosphorylation Site in Cell Transformation and Apoptosis Induced by the Adenovirus Type 5 E1A Protein. *J. Virol.* **70,** 5373–5383 (1996).

154. Whalen, S. G. *et al.* Phosphorylation within the transactivation domain of adenovirus E1A protein by mitogen-activated protein kinase regulates expression of early region 4. *J. Virol.* **71,** 3545–3553 (1997).

155. Zhang, Q., Yao, H., Vo, N. & Goodman, R. H. Acetylation of adenovirus E1A regulates binding of the transcriptional corepressor CtBP. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 14323–14328 (2000).

156. Mercer, S. E. & Friedman, E. Mirk/Dyrk1B: A Multifunctional Dual-Specificity Kinase Involved in Growth Arrest, Differentiation, and Cell Survival. *Cell Biochem. Biophys.* **45,** 303–315 (2006).

157. Liang, Y.-J., Chang, H.-S., Wang, C.-Y. & Yu, W. C. Y. DYRK1A stabilizes HPV16E7 oncoprotein through phosphorylation of the threonine 5 and threonine 7 residues☆. *Int. J. Biochem. Cell Biol.* **40,** 2431–2441 (2008).

158. Firzlaff, J. M., Galloway, D. A., Eisenman, R. N. & Lüscher, B. The E7 protein of human papillomavirus type 16 is phosphorylated by casein kinase II. *New Biol.* **1,** 44–53 (1989).

159. Barbosa, M. S. *et al.* The region of the HPV E7 oncoprotein homologous to adenovirus E1a and Sv40 large T antigen contains separate domains for Rb binding and casein kinase II phosphorylation. *EMBO J.* **9,** 153–160 (1990).

160. Firzlaff, J. M., Lüscher, B. & Eisenman, R. N. Negative charge at the casein kinase II phosphorylation site is important for transformation but not for Rb protein binding by the E7 protein of human papillomavirus type 16. *Proc. Natl. Acad. Sci. U. S. A.* **88,** 5187–5191 (1991).

161. Kim, J. S. *et al.* Protein kinase CK2alpha as an unfavorable prognostic marker and novel therapeutic target in acute myeloid leukemia. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **13,** 1019–1028 (2007).

162. Marin, O. *et al*. Tyrosine versus serine/threonine phosphorylation by protein kinase casein kinase-2. A study with peptide substrates derived from immunophilin Fpr3. *J. Biol. Chem*. **274,** 29260–29265 (1999).

163. Krek, W. Casein kinase II is a predominantly nuclear enzyme. *J. Cell Biol*. **116,** 43–55 (1992).

164. Issinger, O.-G. Casein kinases: pleiotropic mediators of cellular regulation. *Pharmacol. Ther*. **59,** 1–30 (1993).

165. Schn lzer, M. Induced Release of Cell Surface Protein Kinase Yields CK1- and CK2-like Enzymes in Tandem. *J. Biol. Chem*. **271,** 111–119 (1996).

166. Chester, N., Yu, I. J. & Marshak, D. R. Identification and characterization of protein kinase CKII isoforms in HeLa cells. Isoform-specific differences in rates of assembly from catalytic and regulatory subunits. *J. Biol. Chem*. **270,** 7501–7514 (1995).

167. Niefind, K., Guerra, B., Ermakowa, I. & Issinger, O. G. Crystal structure of human protein kinase CK2: insights into basic properties of the CK2 holoenzyme. *EMBO J*. **20,** 5320–5331 (2001).

168. Lozeman, F. J. *et al*. Isolation and characterization of human cDNA clones encoding the alpha and the alpha' subunits of casein kinase II. *Biochemistry (Mosc.)* **29,** 8436–8447 (1990).

169. Glover, C. V. On the physiological role of casein kinase II in Saccharomyces cerevisiae. *Prog. Nucleic Acid Res. Mol. Biol*. **59,** 95–133 (1998).

170. Maridor, G., Park, W., Krek, W. & Nigg, E. A. Casein kinase II. cDNA sequences, developmental expression, and tissue distribution of mRNAs for alpha, alpha', and beta subunits of the chicken enzyme. *J. Biol. Chem*. **266,** 2362–2368 (1991).

171. Litchfield, D. W. *et al*. Functional specialization of CK2 isoforms and characterization of isoform-specific binding partners. *Mol. Cell. Biochem*. **227,** 21–29 (2001).

172. Shi, X. *et al*. A novel casein kinase 2 alpha-subunit regulates membrane protein traffic in the human hepatoma cell line HuH-7. *J. Biol. Chem*. **276,** 2075–2082 (2001).

173. Litchfield, D. W. Protein kinase CK2: structure, regulation and role in cellular decisions of life and death. *Biochem. J.* **369,** 1–15 (2003).

174. Kuenzel, E. A., Mulligan, J. A., Sommercorn, J. & Krebs, E. G. Substrate specificity determinants for casein kinase II as deduced from studies with synthetic peptides. *J. Biol. Chem.* **262,** 9136–9140 (1987).

175. Marchiori, F. *et al.* Synthetic peptide substrates for casein kinase 2. Assessment of minimum structural requirements for phosphorylation. *Biochim. Biophys. Acta* **971,** 332–338 (1988).

176. Perich, J. W., Meggio, F., Reynolds, E. C., Marin, O. & Pinna, L. A. Role of phosphorylated aminoacyl residues in generating atypical consensus sequences which are recognized by casein kinase-2 but not by casein kinase-1. *Biochemistry (Mosc.)* **31,** 5893–5897 (1992).

177. Meggio, F., Marin, O. & Pinna, L. A. Substrate specificity of protein kinase CK2. *Cell. Mol. Biol. Res.* **40,** 401–409 (1994).

178. Sarno, S. *et al.* Mapping the Residues of Protein Kinase CK2 Implicated in Substrate Recognition: Mutagenesis of Conserved Basic Residues in the α-Subunit. *Biochem. Biophys. Res. Commun.* **206,** 171–179 (1995).

179. Sarno, S., Vaglio, P., Meggio, F., Issinger, O. G. & Pinna, L. A. Protein kinase CK2 mutants defective in substrate recognition. Purification and kinetic analysis. *J. Biol. Chem.* **271,** 10595–10601 (1996).

180. Landrieu, I. *et al.* NMR analysis of a Tau phosphorylation pattern. *J. Am. Chem. Soc.* **128,** 3575–3583 (2006).

181. Selenko, P. *et al.* In situ observation of protein phosphorylation by high-resolution NMR spectroscopy. *Nat. Struct. Mol. Biol.* **15,** 321–329 (2008).

182. Theillet, F.-X. *et al.* Cell signaling, post-translational protein modifications and NMR spectroscopy. *J. Biomol. NMR* **54,** 217–236 (2012).

183. Dose, A., Liokatis, S., Theillet, F.-X., Selenko, P. & Schwarzer, D. NMR profiling of histone deacetylase and acetyl-transferase activities in real time. *ACS Chem. Biol.* **6,** 419–424 (2011).

184. Andrew, C. D., Warwicker, J., Jones, G. R. & Doig, A. J. Effect of Phosphorylation on α-Helix Stability as a Function of Position †. *Biochemistry (Mosc.)* **41,** 1897–1905 (2002).

185. Patchell, V. B. *et al*. Phosphorylation of the minimal inhibitory region at the C-terminus of caldesmon alters its structural and actin binding properties. *Biochim. Biophys. Acta BBA - Protein Struct. Mol. Enzymol.* **1596,** 121–130 (2002).

186. Bielska, A. A. & Zondlo, N. J. Hyperphosphorylation of Tau Induces Local Polyproline II Helix †. *Biochemistry (Mosc.)* **45,** 5527–5537 (2006).

187. Tait, S. *et al*. Local control of a disorder–order transition in 4E-BP1 underpins regulation of translation via eIF4E. *Proc. Natl. Acad. Sci.* **107,** 17627–17632 (2010).

188. Sibille, N. *et al*. Structural characterization by nuclear magnetic resonance of the impact of phosphorylation in the proline-rich region of the disordered Tau protein. *Proteins Struct. Funct. Bioinforma.* **80,** 454–462 (2012).

189. Walsh, C. T., Garneau-Tsodikova, S. & Gatto, G. J. Protein posttranslational modifications: the chemistry of proteome diversifications. *Angew. Chem. Int. Ed Engl.* **44,** 7342–7372 (2005).

190. Deribe, Y. L., Pawson, T. & Dikic, I. Post-translational modifications in signal integration. *Nat. Struct. Mol. Biol.* **17,** 666–672 (2010).

191. Venter, J. C. The Sequence of the Human Genome. *Science* **291,** 1304–1351 (2001).

192. Lander, E. S. *et al*. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

193. Cohen, P. The regulation of protein function by multisite phosphorylation--a 25 year update. *Trends Biochem. Sci.* **25,** 596–601 (2000).

194. Yang, X.-J. Multisite protein modification and intramolecular signaling. *Oncogene* **24,** 1653–1662 (2005).

195. Khoury, G. A., Baliban, R. C. & Floudas, C. A. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci. Rep.* **1,** (2011).

196. Hunter, T. Tyrosine phosphorylation: thirty years and counting. *Curr. Opin. Cell Biol.* **21,** 140–146 (2009).

197. Madden, T. L., Tatusov, R. L. & Zhang, J. Applications of network BLAST server. *Methods Enzymol*. **266,** 131–141 (1996).

198. Gasteiger, E. *et al*. in *The Proteomics Protocols Handbook,* 571–607 (Humana Press, 2005).

199. Sigrist, C. J. A. *et al*. PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform*. **3,** 265–274 (2002).

200. Huang, H.-D., Lee, T.-Y., Tzeng, S.-W. & Horng, J.-T. KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Res*. **33,** W226–W229 (2005).

201. Romero, P., Obradovic, Z., Kissinger, C., Villafranca, J. E. & Dunker, A. K. Identifying disordered regions in proteins from amino acid sequence. in **1,** 90–95 (IEEE, 1997).

202. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. Critical assessment of methods of protein structure prediction (CASP)--round x. *Proteins* **82 Suppl 2,** 1–6 (2014).

203. Dosztányi, Z., Csizmók, V., Tompa, P. & Simon, I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol*. **347,** 827–839 (2005).

204. Linding, R., Russell, R. B., Neduva, V. & Gibson, T. J. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res*. **31,** 3701–3708 (2003).

205. Schlessinger, A., Punta, M. & Rost, B. Natively unstructured regions in proteins identified from contact predictions. *Bioinforma. Oxf. Engl*. **23,** 2376–2384 (2007).

206. Prilusky, J. *et al*. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinforma. Oxf. Engl*. **21,** 3435–3438 (2005).

207. Jones, D. T. & Ward, J. J. Prediction of disordered regions in proteins from position specific score matrices. *Proteins* **53 Suppl 6,** 573–578 (2003).

208. Ishida, T. & Kinoshita, K. PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res*. **35,** W460-464 (2007).

209. Huang, Y. J., Acton, T. B. & Montelione, G. T. DisMeta: a meta server for construct design and optimization. *Methods Mol. Biol. Clifton NJ* **1091,** 3–16 (2014).

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

211. Rosano, G. L. & Ceccarelli, E. A. Recombinant protein expression in Escherichia coli: advances and challenges. *Front. Microbiol.* **5,** (2014).

212. Shiloach, J. & Fass, R. Growing E. coli to high cell density--a historical perspective on method development. *Biotechnol. Adv.* **23,** 345–357 (2005).

213. Lee, S. Y. High cell-density culture of Escherichia coli. *Trends Biotechnol.* **14,** 98–105 (1996).

214. Pope, B. & Kent, H. M. High efficiency 5 min transformation of Escherichia coli. *Nucleic Acids Res.* **24,** 536–537 (1996).

215. Bolivar, F., Rodriguez, R. L., Betlach, M. C. & Boyer, H. W. Construction and characterization of new cloning vehicles. I. Ampicillin-resistant derivatives of the plasmid pMB9. *Gene* **2,** 75–93 (1977).

216. Sockolosky, J. T. & Szoka, F. C. Periplasmic production via the pET expression system of soluble, bioactive human growth hormone. *Protein Expr. Purif.* **87,** 129–135 (2013).

217. Mirecka, E. A., Rudolph, R. & Hey, T. Expression and purification of His-tagged HPV16 E7 protein active in pRb binding. *Protein Expr. Purif.* **48,** 281–291 (2006).

218. Calçada, E. O. Intrinsically disordered proteins from sample preparation to molecular basis of function. (International Doctorate in Mechanistic and Structural Systems Biology.University of Florence, IT, 2013).

219. Froger, A. & Hall, J. E. Transformation of Plasmid DNA into E. coli Using the Heat Shock Method. *J. Vis. Exp.* (2007). doi:10.3791/253

220. Hunt, I. From gene to protein: a review of new and enabling technologies for multi-parallel protein expression. *Protein Expr. Purif.* **40,** 1–22 (2005).

221. Lesley, S. A. Parallel methods for expression and purification. *Methods Enzymol.* **463,** 767–785 (2009).

222. Prasad, S., Khadatare, P. B. & Roy, I. Effect of chemical chaperones in improving the solubility of recombinant proteins in Escherichia coli. *Appl. Environ. Microbiol.* **77,** 4603–4609 (2011).

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

224. Wilkins, M. R. *et al.* Protein identification and analysis tools in the ExPASy server. *Methods Mol. Biol. Clifton NJ* **112,** 531–552 (1999).

225. Block, H. *et al.* in *Methods in Enzymology* **463,** 439–473 (Elsevier, 2009).

226. Wong, J. W., Albright, R. L. & Wang, N.-H. L. Immobilized Metal Ion Affinity Chromatography (IMAC) Chemistry and Bioseparation Applications. *Sep. Purif. Rev.* **20,** 49–106 (1991).

227. Gaberc-Porekar, V. & Menart, V. Perspectives of immobilized-metal affinity chromatography. *J. Biochem. Biophys. Methods* **49,** 335–360 (2001).

228. Erickson, H. P. Size and Shape of Protein Molecules at the Nanometer Level Determined by Sedimentation, Gel Filtration, and Electron Microscopy. *Biol. Proced. Online* **11,** 32–51 (2009).

229. Singh, S. M. & Panda, A. K. Solubilization and refolding of bacterial inclusion body proteins. *J. Biosci. Bioeng.* **99,** 303–310 (2005).

230. Fischer, B., Sumner, I. & Goodenough, P. Isolation, renaturation, and formation of disulfide bonds of eukaryotic proteins expressed inEscherichia coli as inclusion bodies. *Biotechnol. Bioeng.* **41,** 3–13 (1993).

231. Dill, K. A. & Shortle, D. Denatured states of proteins. *Annu. Rev. Biochem.* **60,** 795–825 (1991).

232. Tsumoto, K. *et al.* Role of Arginine in Protein Refolding, Solubilization, and Purification. *Biotechnol. Prog.* **20,** 1301–1308 (2004).

233. Samuel, D. *et al.* Proline inhibits aggregation during protein refolding. *Protein Sci. Publ. Protein Soc.* **9,** 344–352 (2000).

234. Cleland, J. L. *et al.* Polyethylene glycol enhanced protein refolding. *Biotechnol. Nat. Publ. Co.* **10,** 1013–1019 (1992).

235. Petersen, S. B., Jonson, V., Fojan, P., Wimmer, R. & Pedersen, S. Sorbitol prevents the self-aggregation of unfolded lysozyme leading to and up to 13 degrees C stabilisation of the folded form. *J. Biotechnol.* **114,** 269–278 (2004).

236. Gekko, K. & Timasheff, S. N. Mechanism of protein stabilization by glycerol: preferential hydration in glycerol-water mixtures. *Biochemistry (Mosc.)* **20,** 4667–4676 (1981).

237. Hagen, A. J., Hatton, T. A. & Wang, D. I. Protein refolding in reversed micelles. *Biotechnol. Bioeng.* **35,** 955–965 (1990).

238. Nara, T. Y. *et al.* Use of zeolite to refold a disulfide-bonded protein. *Colloids Surf. B Biointerfaces* **68,** 68–73 (2009).

239. Dahiya, V. & Chaudhuri, T. K. Chaperones GroEL/GroES Accelerate the Refolding of a Multidomain Protein through Modulating On-pathway Intermediates. *J. Biol. Chem.* **289,** 286–298 (2014).

240. Pardi, A., Wagner, G. & WüThrich, K. Protein conformation and proton nuclear-magnetic-resonance chemical shifts. *Eur. J. Biochem.* **137,** 445–454 (1983).

241. Bax, A. Two-Dimensional NMR and Protein Structure. *Annu. Rev. Biochem.* **58,** 223–256 (1989).

242. Yao, J., Dyson, H. J. & Wright, P. E. Chemical shift dispersion and secondary structure prediction in unfolded and partly folded proteins. *FEBS Lett.* **419,** 285–289 (1997).

243. Sattler, M. Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. Nucl. Magn. Reson. Spectrosc.* **34,** 93–158 (1999).

244. *NMR of proteins and small biomolecules*. (Springer, 2012).

245. Bermel, W., Bertini, I., Felli, I. C., Kümmerle, R. & Pierattelli, R. Novel 13C direct detection experiments, including extension to the third dimension, to perform the complete assignment of proteins. *J. Magn. Reson. San Diego Calif 1997* **178,** 56–64 (2006).

246. Theillet, F.-X. *et al.* The alphabet of intrinsic disorder: I. Act like a Pro: On the abundance and roles of proline residues in intrinsically disordered proteins. *Intrinsically Disord. Proteins* **1,** e24360 (2013).

247. Cornilescu, G., Marquardt, J. L., Ottiger, M. & Bax, A. Validation of Protein Structure from Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase. *J. Am. Chem. Soc.* **120,** 6836–6837 (1998).

248. Cornilescu, G. & Bax, A. Measurement of Proton, Nitrogen, and Carbonyl Chemical Shielding Anisotropies in a Protein Dissolved in a Dilute Liquid Crystalline Phase. *J. Am. Chem. Soc.* **122,** 10143–10154 (2000).

249. Tamiola, K. & Mulder, F. A. A. Using NMR chemical shifts to calculate the propensity for structural order and disorder in proteins. *Biochem. Soc. Trans.* **40,** 1014–1020 (2012).

250. Tamiola, K., Acar, B. & Mulder, F. A. A. Sequence-Specific Random Coil Chemical Shifts of Intrinsically Disordered Proteins. *J. Am. Chem. Soc.* **132,** 18000–18003 (2010).

251. Wang, Y. & Jardetzky, O. Probability-based protein secondary structure identification using combined NMR chemical-shift data. *Protein Sci. Publ. Protein Soc.* **11,** 852–861 (2002).

252. Marsh, J. A., Singh, V. K., Jia, Z. & Forman-Kay, J. D. Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein: implications for fibrillation. *Protein Sci. Publ. Protein Soc.* **15,** 2795–2804 (2006).

253. Solomon, I. Relaxation Processes in a System of Two Spins. *Phys. Rev.* **99,** 559–565 (1955).

254. Kumar, A., Ernst, R. R. & Wüthrich, K. A two-dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton-proton cross-relaxation networks in biological macromolecules. *Biochem. Biophys. Res. Commun.* **95,** 1–6 (1980).

255. Tugarinov, V., Kay, L. E., Ibraghimov, I. & Orekhov, V. Y. High-resolution four-dimensional 1H-13C NOE spectroscopy using methyl-TROSY, sparse data acquisition, and multidimensional decomposition. *J. Am. Chem. Soc.* **127,** 2767–2775 (2005).

256. Clore, G. M. & Gronenborn, A. M. Determination of three-dimensional structures of proteins in solution by nuclear magnetic resonance spectroscopy. *Protein Eng. Des. Sel.* **1,** 275–288 (1987).

257. Tolman, J. R., Flanagan, J. M., Kennedy, M. A. & Prestegard, J. H. Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution. *Proc. Natl. Acad. Sci. U. S. A.* **92,** 9279–9283 (1995).

258. Bertini, I. *et al*. Experimentally exploring the conformational space sampled by domain reorientation in calmodulin. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 6841–6846 (2004).

259. Wöhnert, J., Franz, K. J., Nitz, M., Imperiali, B. & Schwalbe, H. Protein alignment by a coexpressed lanthanide-binding tag for the measurement of residual dipolar couplings. *J. Am. Chem. Soc.* **125,** 13338–13339 (2003).

260. Su, X.-C. & Otting, G. Paramagnetic labelling of proteins and oligonucleotides for NMR. *J. Biomol. NMR* **46,** 101–112 (2010).

261. Sass, H. J., Musco, G., Stahl, S. J., Wingfield, P. T. & Grzesiek, S. Solution NMR of proteins within polyacrylamide gels: diffusional properties and residual alignment by mechanical stress or embedding of oriented purple membranes. *J. Biomol. NMR* **18,** 303–309 (2000).

262. Tycko, R., Blanco, F. J. & Ishii, Y. Alignment of Biopolymers in Strained Gels: A New Way To Create Detectable Dipole–Dipole Couplings in High-Resolution Biomolecular NMR. *J. Am. Chem. Soc.* **122,** 9340–9341 (2000).

263. Hansen, M. R., Mueller, L. & Pardi, A. Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions. *Nat. Struct. Biol.* **5,** 1065–1074 (1998).

264. Clore, G. M., Starich, M. R. & Gronenborn, A. M. Measurement of Residual Dipolar Couplings of Macromolecules Aligned in the Nematic Phase of a Colloidal Suspension of Rod-Shaped Viruses. *J. Am. Chem. Soc.* **120,** 10571–10572 (1998).

265. Tjandra, N. & Bax, A. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science* **278,** 1111–1114 (1997).

266. Rückert, M. & Otting, G. Alignment of Biological Macromolecules in Novel Nonionic Liquid Crystalline Media for NMR Experiments. *J. Am. Chem. Soc.* **122,** 7793–7797 (2000).

267. Sanders, C. R. & Schwonek, J. P. Characterization of magnetically orientable bilayers in mixtures of dihexanoylphosphatidylcholine and dimyristoylphosphatidylcholine by solid-state NMR. *Biochemistry (Mosc.)* **31,** 8898–8905 (1992).

268. Harroun, T. A. *et al*. Comprehensive examination of mesophases formed by DMPC and DHPC mixtures. *Langmuir ACS J. Surf. Colloids* **21,** 5356–5361 (2005).

269. Katsaras, J., Harroun, T. A., Pencer, J. & Nieh, M.-P. 'Bicellar' lipid mixtures as used in biochemical and biophysical studies. *Naturwissenschaften* **92,** 355–366 (2005).

270. van Dam, L., Karlsson, G. & Edwards, K. Direct observation and characterization of DMPC/DHPC aggregates under conditions relevant for biological solution NMR. *Biochim. Biophys. Acta* **1664,** 241–256 (2004).

271. Nieh, M.-P. *et al*. Magnetically alignable phase of phospholipid 'bicelle' mixtures is a chiral nematic made up of wormlike micelles. *Langmuir ACS J. Surf. Colloids* **20,** 7893–7897 (2004).

272. Ottiger, M. & Bax, A. Characterization of magnetically oriented phospholipid micelles for measurement of dipolar couplings in macromolecules. *J. Biomol. NMR* **12,** 361–372 (1998).

273. Cavagnero, S., Dyson, H. J. & Wright, P. E. Improved low pH bicelle system for orienting macromolecules over a wide temperature range. *J. Biomol. NMR* **13,** 387–391 (1999).

274. Liu, P.-H. & Ding, S. Application of Weakly Oriented Non-ionic Liquid Crystal Solvents to NMR Studies of Poly-ethylene Oxide (PEO). *J. Chin. Chem. Soc.* **50,** 201–204 (2003).

275. Tjandra, N. Establishing a degree of order: obtaining high-resolution NMR structures from molecular alignment. *Structure* **7,** R205–R211 (1999).

276. Bax, A. Weak alignment offers new NMR opportunities to study protein structure and dynamics. *Protein Sci. Publ. Protein Soc.* **12,** 1–16 (2003).

277. Palmer, A. G. Probing molecular motion by NMR. *Curr. Opin. Struct. Biol.* **7,** 732–737 (1997).

278. Kleckner, I. R. & Foster, M. P. An introduction to NMR-based approaches for measuring protein dynamics. *Biochim. Biophys. Acta BBA - Proteins Proteomics* **1814,** 942–968 (2011).

279. Jarymowycz, V. A. & Stone, M. J. Fast time scale dynamics of protein backbones: NMR relaxation methods, applications, and functional consequences. *Chem. Rev.* **106,** 1624–1671 (2006).

280. Peng, J. W. & Wagner, G. Investigation of protein motions via relaxation measurements. *Methods Enzymol.* **239,** 563–596 (1994).

281. Kay, L. E., Torchia, D. A. & Bax, A. Backbone dynamics of proteins as studied by 15N inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. *Biochemistry (Mosc.)* **28,** 8972–8979 (1989).

282. Camilloni, C., Cavalli, A. & Vendruscolo, M. Replica-Averaged Metadynamics. *J. Chem. Theory Comput.* **9,** 5610–5617 (2013).

283. Fu, B. & Vendruscolo, M. in *Intrinsically Disordered Proteins Studied by NMR Spectroscopy* (eds. Felli, I. C. & Pierattelli, R.) **870,** 35–48 (Springer International Publishing, 2015).

284. Ozenne, V. *et al.* Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics* **28,** 1463–1470 (2012).

285. Blanquet, P. R. Casein kinase 2 as a potentially important enzyme in the nervous system. *Prog. Neurobiol.* **60,** 211–246 (2000).

286. Hathaway, G. M. & Traugh, J. A. Interaction of polyamines and magnesium with casein kinase II. *Arch. Biochem. Biophys.* **233,** 133–138 (1984).

287. Hošek, T. Development of new NMR methods to study intrinsically disordered proteins. (2014).

288. Hošek, T. *et al.* Structural and Dynamic Characterization of the Molecular Hub Early Region 1A (E1A) from Human Adenovirus. *Chem. - Eur. J.* **22,** 13010–13013 (2016).

289. Felli, I. C. & Pierattelli, R. Novel methods based on (13)C detection to study intrinsically disordered proteins. *J. Magn. Reson. San Diego Calif 1997* **241,** 115–125 (2014).

290. Brutscher, B. *et al*. NMR Methods for the Study of Instrinsically Disordered Proteins Structure, Dynamics, and Interactions: General Overview and Practical Guidelines. *Adv. Exp. Med. Biol.* **870,** 49–122 (2015).

.