UNIVERSITÀ
DEGLI STUDI
FIRENZE

DOTTORATO DI RICERCA
INTERNATIONAL DOCTORATE IN MECHANISTIC AND STRUCTURAL SYSTEMS BIOLOGY

CYCLE XXVI

COORDINATOR Prof. Claudio Luchinat

INTRINSICALLY DISORDERED PROTEINS
FROM SAMPLE PREPARATION TO MOLECULAR BASIS OF FUNCTION

S.S.D. CHIM/03

Candidate
Dr. Eduardo O. Calçada

Tutor
Prof. Roberta Pierattelli

Coordinator
Prof. Claudio Luchinat

Year 2011/2013

This thesis has been approved by the University of Florence, the University of Frankfurt and the Utrecht University

Universiteit Utrecht

CERM Firenze

FRANKFURT AM MAIN

Abstract

Intrinsically disordered proteins (IDPs) are characterized by the lack of a well-structured three-dimensional fold. This property confers them functional advantages, complementary to those of folded proteins, which need to be characterized to expand our view of how protein structural and dynamic features affect function beyond the static picture of a single well defined 3D structure.

IDPs are often able to bind to different partners or to act as hub proteins, playing in this way an important role in a variety of different processes. IDPs are particularly relevant for viruses which need to exploit simple amino acid sequences (short linear motifs, SLiMs), well exposed and ready to function, to interact with crucial key proteins from the host organism. However, the molecular basis concerning the mechanisms of these interactions are not yet completely understood.

In this work we present the study of two viral proteins, the E7 from Human Papilloma Virus and the E1A from Human Adenovirus. E7 and E1A share sequence and structural similarities, shown to be crucial for interaction and deregulation of key proteins from the host, towards cancer development. Here we will show the structural and dynamic characterization of these two proteins by Nuclear Magnetic Resonance (NMR), supported by other biophysical techniques like Small-Angle X-ray Scattering (SAXS), laying the basis for the study of their interactions with medically relevant targets. Besides the importance for evolutionary biology, understanding the molecular basis of these interactions will open a wide range of opportunities to understand dozens of interactions of these viral IDPs with their partners through their short linear motifs.

Keywords

IDPs, SLiMs, HPV16 E7, HAdV5 E1A, CBP/p300, ID4, NMR, SAXS

Acknowledgements

The present thesis is the result of a long travel. I had the opportunity to grow not only as a scientist but also as a person. Joining a Marie Curie program made me get in contact with an international network of research institutes and principal investigators leaders in diverse research fields. After this period I've got a feeling that more than a PhD title I earned a huge knowledge in many different fields spanning from biophysics to molecular biology and biochemistry. I had also the opportunity to teach, which I quite appreciated, co-supervising the BSc thesis work of four students, as well the MSc and PhD work of other two students. Working in CERM, Italy gave me also the opportunity to meet Ivano Bertini, who left a giant footprint on metalloprotein research through NMR. I want to remind his words when I met him few years ago, and I introduced myself saying I knew nothing about NMR: "...you know, a good molecular biologist will be always the limiting step for of any good 'NMRist' doing structural biology..." it motivated me!

I want to acknowledge my PhD supervisor Roberta Pierattelli and co-supervisor Isabella Felli for their support and transmission of knowledge and for having always trusted on me. Many thanks to my friend and lab colleague Tomáš Hošek for all the help and patience, and even more for the funny moments shared during this course. During my PhD I had the opportunity to personally met Peter Tompa, Vladimir Uversky and Keith Dunker who made a big contribution to glimpse the new world of intrinsically disordered proteins. I'm glad also to have the opportunity to meet and had a long chat in an informal way with the 1991 Nobel Prize in Chemistry Richard R. Ernest.

I want to acknowledge the support from my fellows in Portugal; special thanks to Ricardo Louro who has been always a great leader. Many thanks to my family that helped making the distance shorter, Irene, José, Mauro and Nelson; and my friends Angela, Aníbal, Isabel, Márcia, Ricardo, Rodrigo and Sérgio. I wish my parents were present watching me entering in this journey and reaching the end; I'm sure they would be thrilled.

Of course nothing of this was possible without the unconditional support of my wife Adelaide who has been always present on the most important moments of my life, and recently bringing to us our lovely daughter Leonor, the most important chapter of my PhD. This thesis is dedicated to her.

Table of contents

List of Abbreviations

AdV	Adenovirus
CBP	CREB-Binding Protein
DTT	Dithiothreitol
DLS	Dynamic Light Scattering
E2F	E2 Transcription Factor
HAdV	Human Adenovirus
HMQC	Heteronuclear multiple quantum coherence spectroscopy
HPV	Human Papilloma Virus
HSQC	Heteronuclear single quantum coherence spectroscopy
IDP	Intrinsically Disordered Protein
IDPR	Intrinsically Disordered Protein Region
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effect
pRB/RB	Retinoblastoma Tumor Suppressor Protein pocket
SAXS	Small-Angle X-ray Scattering
SLiMs	Short Linear Motifs

1. Introduction

The so-called structure-function paradigm states that a three-dimensional structure is a prerequisite for a protein to be functional (Wright and Dyson, 1999). Despite initial reports showing the protein-function relationship present in an unstructured state, reviewed in (Uversky, 2013a), the occasional discovery of functional disordered proteins was largely overshadowed by the regular appearance of specific three-dimensional protein structures obtained through X-ray crystallography or NMR spectroscopy methods (Rezaei Ghaleh *et al*., 2012). Only in the '90s the importance of local flexibility in determining the function of proteins has been recognized and widely scrutinized (Wagner, 1993).

Comparison of the results of the X-ray and NMR analyses of the same protein has revealed that solution and crystal structures can be quite different, with solution structures being, is some cases, much more flexible (Lian, 1998) than those obtained by crystallography. In several cases NMR spectroscopy was able to describe protein segments that were missing in crystallographic experiments and to characterize flexible regions that were escaping detection or had become rigid due to crystal packing (Mittermaier and Kay, 2006). Proteins lacking a stable well defined 3D structure characterized by extensive flexibility have been named in the literature in many different ways like "intrinsically unstructured", "natively denatured", "natively unfolded" among others (Dunker *et al*., 2013). Intrinsically disordered proteins (IDPs) has now become the widely accepted term to indicate these proteins. "Intrinsically" rather than "natively" as many IDPs can fold upon binding to the partner, and "disordered" rather than "unstructured" to highlight the possible transient existence of local and global structures (Rezaei Ghaleh *et al*., 2012). In this way intrinsically disordered regions (IDRs) present in mainly folded proteins can also be defined. The Disprot database [http ://www.disprot.org, (Sickmeier *et al*., 2007)] latest release (May 2013) contains 1539 disordered regions in 694 IDPs, numbers that are increasing over the years (Figure 1.1) even if IDPs/IDRs are often difficult to categorize. Recently a new database of IDPs was created to combine all the information obtained from different biophysical techniques such as NMR and SAXS, the Protein Ensemble Data Base (pE-DB, http://pedb.vib.be) where IDPs are described as structural ensembles (Varadi *et al*., 2013).

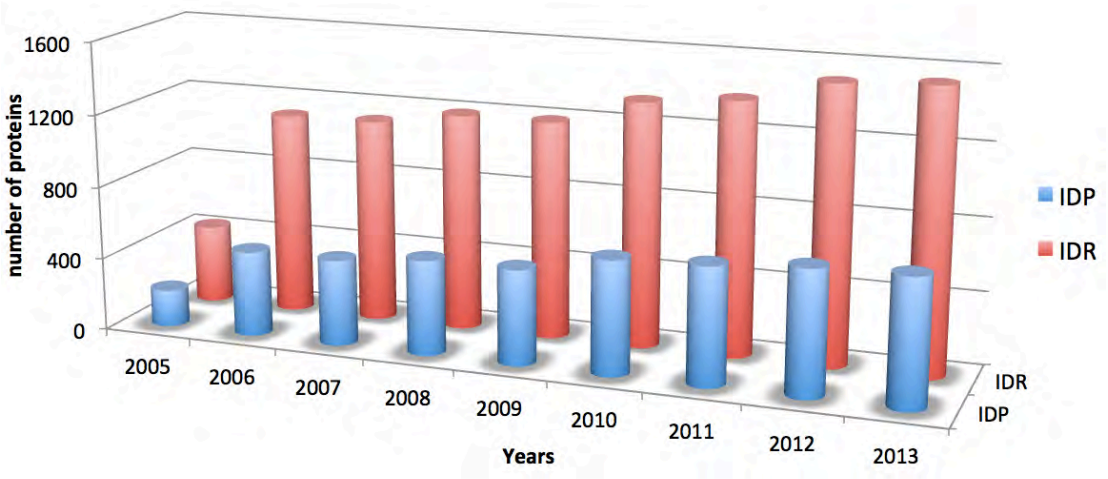


Figure 1.1. IDPRs and IDPs reported on Disprot database, since its start in 2005.

If the extent of local flexibility is taken to its extreme conditions it leads to completely random coil behavior of a polypeptide chain (Bertini *et al*., 2012; He *et al*., 2009) through a wide variety of intermediate cases both in terms of extent of mobility or in terms of protein stretches involved, as a continuous spectrum of possible protein states (Uversky, 2013b).

Despite the absence of a rigid secondary or tertiary structure, IDPs fulfill a number of biological functions, that range from regulation of transcription and translation, storage of small molecules and regulation of the self-assembly of large multi-protein complexes (Dunker *et al*., 2008b; Tompa, 2012). Indeed, many examples of IDPs appeared in the literature showing how their structural plasticity, adaptability, intrinsic flexibility, can actually be a key feature to enable them to interact with a variety of different partners, to adapt to different conditions and to act as hub proteins (Mittag and Forman-Kay, 2007; Uversky *et al*., 2005; Wright and Dyson, 1999). These peculiar properties provide functional advantages to IDPs enabling them to play key roles in many regulatory processes (Fuxreiter *et al*., 2008). Thus, IDPs have been recognized as relevant players in pathologies associated with protein misfolding and aggregation, as observed in several human degenerative diseases (Uversky *et al*., 2008). The obsolete perception that IDRs simply act as domain linkers has been replaced by a model where these

regions may contain numerous regulatory and signaling modules (Diella *et al*., 2008; Dyson and Wright, 2005; Fuxreiter *et al*., 2007). Many eukaryotic hub proteins, such as p53, contain large segments of IDRs maximizing interaction interfaces in a constrained space (Haynes *et al*., 2006).

Thanks to the largely extended and exposed backbones, IDPs often display short amino acid segments (short linear motives, SLiMs) that may interact with different partners (Davey *et al*., 2011; Tompa and Fuxreiter, 2008). This strategy of exploiting proteins' SLiMs for driving interactions seems to be extensively used by viruses to infect healthy cells. In virtue of their small genomes, which are able to code only a limited number of proteins, viruses need effective ways to interfere with the host (Davey *et al*., 2011). This is the case of human papilloma virus (HPV) (Chemes *et al*., 2012) as well the human adenovirus (HAdV) (Pelka *et al*., 2008). Several proteins encoded by these two viral genomes play similar roles in several human pathways. Good examples are HPV E7 and HAdV E1A proteins, both playing a crucial interaction with the pRB tumor suppressor, which may lead to cancer development upon viral infection. Moreover, both E7 and E1A interact with key host regulators such as the general transcriptional co-activator CREB binding protein (CBP), a large human protein comprising many globular domains linked by five intrinsically disordered (ID) domains which have not been studied yet (Bernat *et al*., 2003; Ferreon *et al*., 2013).

The HPV E7 protein has been studied for more than 30 years, and its key role on cervix cancer development is established (McLaughlin-Drubin *et al*., 2012). It has been reported that a multitude of different partners interact with E7 but the molecular basis of these interactions at atomic resolution level are not known. The tendency of E7 protein to form oligomers has been also reported that but is not clear how this phenomenon occurs. The high-resolution characterization of the E7 protein is crucial for further studies of protein interaction, oligomer formation and drug design.

The HAdV E1A protein belongs to a different viral origin but presents a high homology with the E7 protein, also interacting with pRB and CBP/p300 (Ferreon *et al*., 2009). The atomic resolution characterization of the full-construct of the E1A protein has never been achieved, compromising the understanding of the molecular basis of interactions with possible partners

like the CBP. Moreover, the comparison between E7 and E1A homology functions has been attempted (Fera and Marmorstein, 2012) but never combining of low and high methods.

Understanding the molecular basis of the functional role of IDPs requires their characterization at atomic resolution. The presence of dynamic ensembles of conformations under physiological conditions makes their analysis difficult to achieve but NMR has been proven a useful tool to analyze IDPs (Bermel *et al*., 2009; Davey *et al*., 2011; Felli and Pierattelli, 2012; Mittag and Forman-Kay, 2007; Wright and Dyson, 1999) in conjunction with low-resolution techniques like circular dichroism (CD), dynamic light scattering (DLS) and small-angle X-ray scattering (SAXS).

Because IDPs are characterized by a continuous spectrum of possible disordered states (Uversky, 2013b), the expression and purification of these complex proteins may represent a serious challenge with respect to well-folded proteins (Tompa, 2012). The NMR characterization of IDPs requires the production of high-purity isotopically labeled samples in sufficient amount for analysis. This involves the use of different techniques for the expression and purification of these particular proteins. Different protocols of expression and purification should be tested in order to optimize the expression systems.

In the following sections the results obtained for both E7 full-length, for a shorter construct E7 CR3 domain as well for both E1A 12S and E1A 13S proteins are presented and discussed. The high-resolution characterization of these proteins opens the way for the understanding the molecular basis through which these viral IDPs hijack cell regulation.

2. The Systems

2.1. Human Papilloma Virus E7 protein

Human papillomaviruses (HPVs) are small double stranded DNA viruses that infect the squamous epithelium, the superficial layer of animal tissue. HPV is a family of more than 100 different viruses (Bernard *et al*., 2010). They can be clinically classified as "low-risk" (LR), causing benign warts, but among the genital HPVs, which are sexually transmitted, 15 are categorized as "high risk" (HR) and are considered the causative agents of most cervical cancers (Moody and Laimins, 2010a), the second most common cancer in women worldwide (Hausen, 2002). Infection with HPV forces quiescent cells to enter the cell division cycle and thus the abnormal cell proliferation initiates progression to the cancer (McLaughlin-Drubin and Münger, 2009; Schiffman *et al*., 2007). The entire genome of HPV includes six nonstructural proteins (E1, E2, E4, E5, E6, and E7) and two structural proteins (L1 and L2). The interactome of all these proteins have been recently reviewed (Xue *et al*., 2013). The HR HPV 16 is considered the most carcinogenic (Hausen, 2002) and its oncogenic potential relies on the expression of two proteins, E6 and E7 (Boulet *et al*., 2007; Narisawa Saito and Kiyono, 2007) which are consistently present in HPV-associated carcinomas and are necessary for induction and maintenance of the transformed phenotype (Münger *et al*., 2004). E6 promotes tumorigenesis by stimulating cellular degradation of the tumor suppressor p53 while E7 binds pRB, the product of the retinoblastoma tumor susceptibility locus, which is an important regulator of cell growth (Malloy and Carson, 2008) (Banks *et al*., 2012). The HPV 16 E6 and E7 genes together are necessary and sufficient for transformation of primary human keratinocytes, the predominant cell type in the epidermis (Münger *et al*., 1989).

E7 is a small protein constituted by about one hundred amino acids. Three conserved regions (CR), namely CR1 and CR2 in the N-terminal half and CR3 in the C-terminal half (Phelps *et al*., 1992) (Todorovic *et al*., 2012) can be identified (Figure 2.1). The first two (CR1 and CR2), which have been found to interact with a high variety of target proteins, (McLaughlin-Drubin and Münger, 2009) show a high propensity to be disordered (Uversky *et al*., 2006). The CR3 region, which contains two Cys-X-X-Cys motifs separated by 29 to 30 amino acids, has been shown to bind zinc (Barbosa *et al*., 1989) and is

expected to be more structured (Uversky *et al*., 2006). The structural characterization of CR3 of the protein, that contains the zinc binding motifs, has been achieved for short constructs comprising only CR3 for the LR HPV 1a variant through X-ray (Liu *et al*., 2006) and for the HR HPV 45 variant through nuclear magnetic resonance (NMR) spectroscopy (Ohlenschläger *et al*, 2006). However the whole protein failed to give crystals and its characterization through NMR was never reported. Thus, atomic resolution information on HPV 16 E7 was lacking despite the interest in accessing high-resolution information on the protein responsible for the high oncogenic potential of this variant, and on its interactions. Sedimentation equilibrium experiments have shown that E7 presents oligomerization properties and can be found as monomers, dimers, and tetramers, but under physiological conditions it is primarily dimeric (Clements *et al*., 2000).

Among the many documented interactions of E7 with host's proteins, (Chemes *et al*., 2012) the one with pRB plays a key role in its oncogenic potential and has been extensively characterized through many techniques (Uversky and Longhi, 2011). The structure of a short polypeptide containing the LXCXE motif localized in the CR2 region (Dyson *et al*., 1989) bound to pRB is available, providing precious hints to understand how this short fragment binds to pRB (Lee *et al*., 1998). However this polypeptide has much lower affinity for pRB respect to the whole protein (Dyson *et al*., 1989) indicating that other parts of the protein also contribute to the binding (Chemes *et al*., 2010). The E7 mutational studies that provided the available functional information on all the three different CRs have been recently reviewed (Roman and Munger, 2013) summarizing almost 30 years of studies on papillomavirus E7 proteins. E7 binds to several other partners, increasing their activity, to enhance pRB degradation. The cullin 2 ubiquitin ligase complex, is bound by the CR1 of the E7 protein (Huh *et al*., 2007).

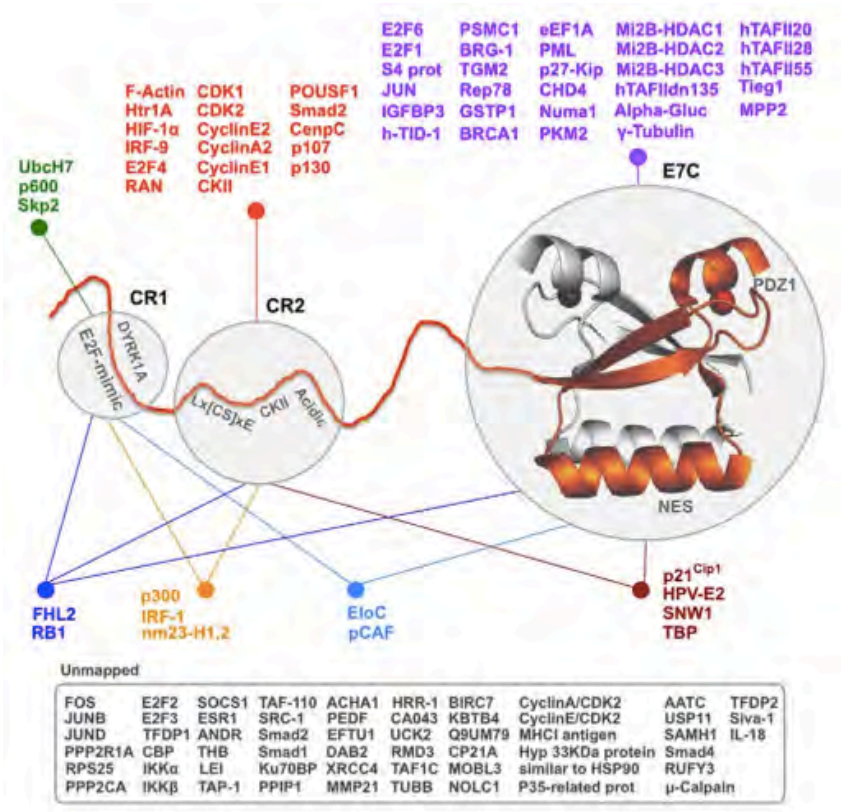


Figure 2.1. Schematic representation of HPV16 E7 oncoprotein including the amino-acid sequence, its division into the three conserved regions and interaction partners (Based on (Chemes *et al*., 2012). CR1 (15 residues), CR2 (22 residues) and CR3 (61 residues) based on HPV45 NMR solution structure (Ohlenschläger *et al*, 2006).

The CR3 also interacts with the subunit 4 (S4) ATPase of the S26 proteasome and enhances its APTase activity providing another pathway for pRB degradation (Berezutskaya *et al*., 1997). E7 interacts through the N-terminal half also with other pRB related tumor suppressor proteins, p107 and 130 (Davies *et al*., 1993; Dyson *et al*., 1992b). E7 also interacts with CBP/p300 proteins family a key host transcriptional coactivator (Bernat *et al*., 2003). These are just a few examples of the complex array of interactions established by viruses like HPV to hijack cell regulation and ensure survival.

To contribute to the understanding of the sophisticated mechanisms behind HPV-related diseases it would be of primary importance to access high resolution structural and dynamic information of its small, though complex, oncogenic proteins.

2.2. Human Adenovirus E1A protein

Adenoviruses (AdV) are medium-size viruses representing the largest non-enveloped viruses. Named for their discovery in adenoid tissue, Human Adenoviruses (HAdV) cause a variety of respiratory, gastrointestinal tract or eye infections. All HAdV are 90nm diameter, with an icosahedral nucleocapsid and non-enveloped particles with fibers protruding from each vertex of the virion (Stewart *et al.*, 1991) (Figure 2.2).

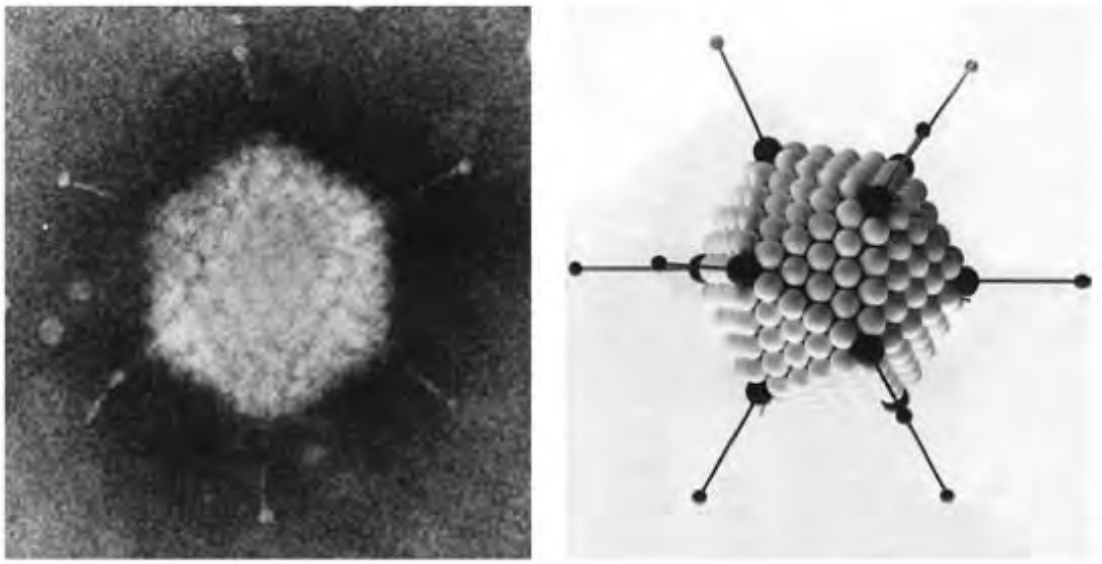


Figure 2.2. The structure of adenoviruses. (a) Electron micrograph of an adenovirus. (b) Model of an adenovirus to show arrangements of the capsomers (Dimmock *et al.*, 2007).

Currently 57 accepted HAdV serotypes are known to transform cultured rodent cells promoting tumorigenesis in immune-compromised rodents (Gallimore, 1972) establishing HAdVs as useful model systems for studying the mechanisms of cancer. HAdVs have played a large role understanding the DNA replication, cell cycle control, transcription, apoptosis, immunological responses to viral infections, and mRNA processing like the importance of mRNA splicing. (Dimmock *et al.*, 2007). These studies of adenovirus oncogenicity demonstrate how important the immune system is in preventing the growth of tumors. In the last few years, the activities of virus-encoded proteins leading to transformation and tumorigenesis have become much clear. The first genomes to be completely sequenced were the closely related

HAdV-2 and 5 and since then they have been extensively manipulated to create viruses that express proteins containing one or more mutations. The genome of HAdV contains approximately 36000 bp of linear, double-stranded DNA encoding five early (E1A, E1B, E2 early, E3, E4), three delayed-early (IX, IVa2, E2 late) which are transcribed by cellular RNA polymerase II (Halldén, 2012). Each of these transcripts is alternatively spliced, giving rise to multiple mRNAs. Moreover, many alternatively spliced products of an individual transcriptional unit encode proteins with similar function, as the case of the two major E1A proteins forms – 13S and 12S. Detailed study of the E1A proteins has been crucial to understand how transcription can be controlled in eukaryotes.

The HAdV E1A protein is the first protein produced post-HAdV infection, and serves two main functions. The first is to modulate host and viral transcription. The second is to force host cell cycle progression into S phase, to promote an optimal environment for viral replication. Early HAdV gene products are responsible for activating transcription of viral genes, inducing host cell cycle progression, initiating viral DNA replication, and blocking apoptosis and host anti-viral responses. The E1A transcriptional unit contains a constitutively active enhancer, and is the first viral gene expressed during infection and is essential for viral replication (Jones and Shenk, 1979b; Nevins *et al*., 1979). E1A gene products are the primary trans-activators of the viral genome, and thus the functions of other early genes intimately depend on E1A (Hearing and Shenk, 1983). The various E1A proteins modulate viral and cellular transcription, induce host cell cycle progression to S phase to create an optimal environment for viral replication, and inhibit host anti-viral responses through a variety of intricate mechanisms. Since E1A proteins have no enzymatic or DNA binding ability, their activities rely on manipulation of cellular factors to exert their effects (Ferguson *et al*., 1985). The E1A gene encodes five proteins, termed 9S through 13S based on the sedimentation coefficients of their respective alternatively spliced mRNAs. 13S (289 residues) and 12S (243 residues) are the main products expressed early during infection, and are the best characterized (Pelka *et al*., 2011; Perricaudet *et al*., 1979) even if no structural information is available. They differ only by a 46 amino acid sequence found in 13S but not 12S. The 11S, 10S, and primarily 9S

are expressed later in infection, however their functions are not well understood (Stephens and Harlow, 1987). It was shown recently that 9S E1A could independently activate expression of viral genes, and promote HAdV replication, and that replication was dependent on an interaction of 9S with the S8 component of the proteasome (Miller *et al.*, 2012). This is intriguing as only 28 amino acids are common between 9S and the larger forms of E1A.

Amino acid sequence alignment of 13S E1A from various serotypes reveals the presence of four main conserved regions (CR1-CR4), which are separated by less conserved regions (Kimelman *et al.*, 1985) as shown on Figure 2.3. Not only the CRs but also other areas of the E1A protein participate in important protein-protein interactions with over 50 cellular factors through SLiMs (Davey *et al.*, 2011; Pelka *et al.*, 2008). The identification of these motifs has been important as a molecular tool as already shown for E7 protein (Chemes *et al.*, 2012).

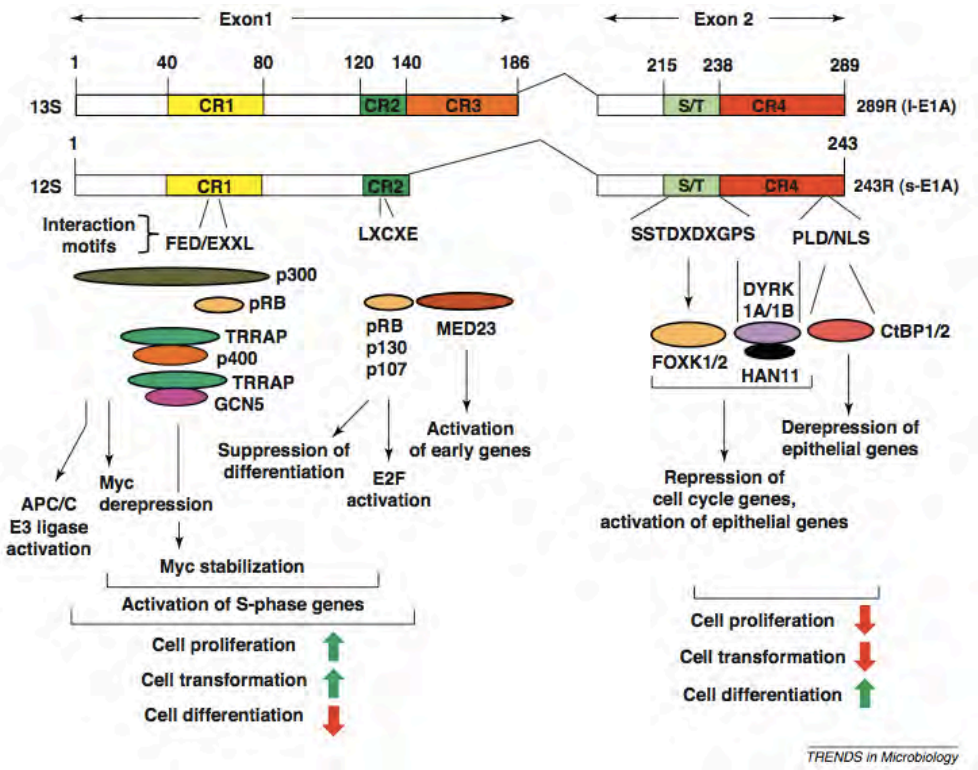


Figure 2.3. Diagram of E1A splice 12S and 13S E1A proteins. B) Alignment of 13S (289R), 12S (243R) proteins and location of conserved regions (CR) 1-4 including the deleted CR3 region from splicing on 12S E1A. The sequence of each HAdV-5 E1A SLiM is indicated bellow the respective CR where the partners interact. The known partners are listed bellow the HAdV-5 SLiM. Adapted from (Chinnadurai, 2011).

Both 13S and 12S E1A can bind pRb (Rb "pocket" domain) through CR1, and more importantly, through an LXCXE motif in CR2. This motif binds to the pRB to displace E2F, bypassing the normal cell cycle regulatory signals, and thereby forcing cell entry into S-phase due to constitutive activation of E2F-responsive promoters (Bandara and La Thangue, 1991; Zamanian and La Thangue, 1992). Adenovirus E1A makes two distinct contacts with the retinoblastoma protein (Dyson *et al*., 1992a). E1A mutants unable to bind pRb still retain the ability to induce an S-phase transition in host cells, through the N-terminus/CR-1, and therefore other known and unknown targets of E1A must be influencing host cell cycling (Zerler *et al*., 1987). The E1A N-terminus (residues 1-41) and CR-3 together are required for optimal early gene transactivation (Wong and Ziff, 1994). They have been shown to interact with repressors such as the p400 chromatin remodeling complex (Fuchs *et al*., 2001) as well with a variety of transcription-activators including histone acetylases p300/CBP and p300/CBP-associated factor (pCAF) (Frisch and Mymryk, 2002; Pelka *et al*., 2008). The interaction of a small fragment of E1A [53-91] with TAZ2 domain from CBP/p300 was shown by NMR (Ferreon *et al*., 2009), opening the way to many possible studies at atomic resolution level.

The CR3 region - the extra 46 residues unique to 13S E1A – contains four cysteines predicted to form a zinc-finger domain, although the structure is still undetermined (Culp and Webster, 1988). HAdVs with complete knockout of 13S E1A or mutations in CR3, experience drastically reduced rates of early gene transcription (Berk *et al*., 1979; Jones and Shenk, 1979a). This is because CR3 functions as a potent trans-activator by binding the DNA-binding domains of transcription factors at early HAdV promoters through its CR3 C-terminal 10 amino acids, and recruiting transcriptional regulators through its N-terminal zinc-finger domain (Lillie and Green, 1989) (Liu and Green, 1994). CR3 interacts not only with transcriptional activators, such as TATA binding protein (TBP), and the p300/CREB-binding protein (CBP) (Pelka *et al*, 2009), but also with repressors such as BS69 (Hateboer *et al*., 1995) and GCN5 (Ablack *et al*, 2012), indicating that CR3 modulation of early gene transcription is complex (Pelka *et al*., 2009). CR3 was also found to interact with CtBP causing hyper-activation of CtBP occupied promoters (Bruton *et al*.,

2008). All this reported interactions of E1A CR3 have been compared across different HAdV subgroups (Ablack *et al*., 2010).

CR4 is the region encoded by the second exon of E1A spanning residues 240 to 288 of HAdV5 E1A (Pelka *et al*., 2008). While E1A can interact with over 50 known proteins, only four binding partners have been identified for CR4. The first E1A CR4 binding partners, and hence the best characterized is the transcriptional co-repressors CtBP (the E1A C-terminus binding protein). Interaction of CtBP with HAdV5 E1A requires the PLDLS SLiM, and this sequence, or closely related variants, is present in all known E1A sequences as well as in those of many cellular proteins that interact with CtBP. This interaction removes CtBP from sequence specific transcription factors, relieving repression of cellular genes (Chinnadurai, 2002). Other partners are the Forkhead transcription factors FOXK1 and FOXK2, which interactions occur outside CR4 and requires residue S219, which must be phosphorylated for binding. The tyrosine-regulated kinase (DYRK) proteins are also reported to interact with CR4 as well the importin-α3 (also known as Qip1) (Cohen *et al*., 2013; Subramanian *et al*., 2013).

Among the many documented interactions of E1A with a variety of different proteins by host cells, one of them attracted our attention, that is the one with the human CREB-binding protein (CBP) (UniProt Q92793). CBP is a co-activator human protein related with long-term memory formation as well with acute amyloid leukemia. Together with p300 (a less extended analogous of CBP protein) CBP/p300 are a family of proteins playing a key role on interaction with G protein signaling. CBP is 2442 residues protein where five IDRs have never been described. A short fragment of this protein named CBP-ID4, constituted by a 202 residues region [1856-2058] and predicted to be completely unfolded (Figure 2.4), is expected to interact with E1A.

However, the interaction was proved between a short E1A fragment [53-91], with the CBP-TAZ2 structured domain (91 residues) that is located before CBP-ID4 (Figure 2.5).

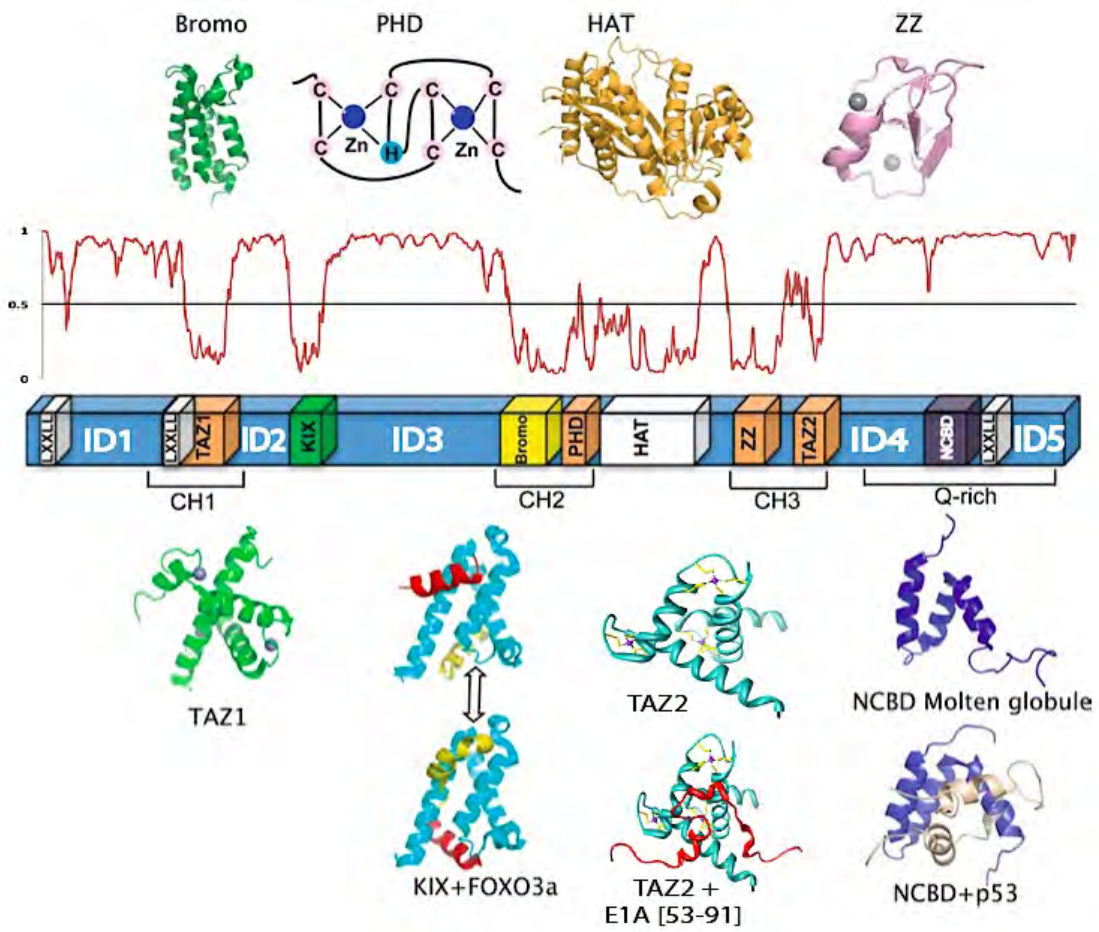


Figure 2.4. Schematic representation of the domain architecture of CBP/p300. The disordered and ordered regions were predicted by PONDR-FIT (Xue *et al.*, 2010). The structures of each domain are shown and labeled. *Top* the CBP Bromo domain (PDB: 3DwY); the zinc-binding mode of the PHD domain of CBP/p300 (no high-resolution structure available); the p300 HAT domain in complex with an inhibitor (PDB: 3BIY); the CBP ZZ domain (PDB: 1TOT). *Bottom* the CBP TAZ1 domain (PDB: 1U2N); KIX domain in complex with FOXO3a PDB: 2LQH and 2LQI); the p300 TAZ2 domain (PDB: 3IO2) and its interaction with E1A CR1 comprising residues [53-91] (PDB: 2KJE); the molten globule state structure of CBP NCBD domain (PDB: 2KKJ), and the structured complex of NCBD with p53 formed through binding-coupled folding (PDB: 2L14). Transactivation domains that primarily interact with transcriptional activators are shown below and those that interact primarily with chromatin are shown on top. Adapted from (Wang *et al.*, 2013).

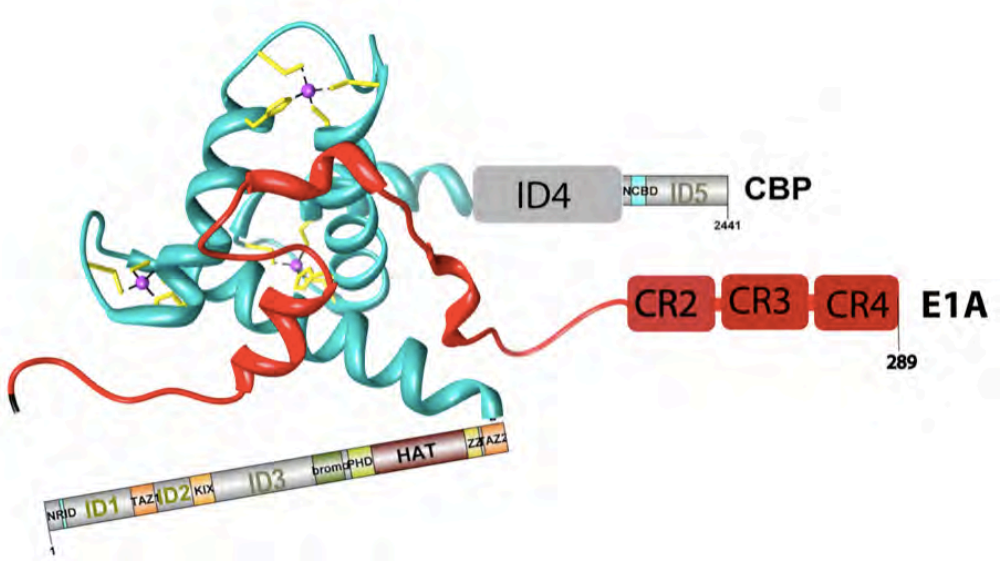


Figure 2.5. Schematic representation of CBP (blue) and E1A (red) interaction. The position of ID4 right after the TAZ2 domain that interacts with E1A CR1 [53-91] may lead to interaction with other regions of E1A due to its proximity in the architecture of CBP.

The analysis of the interaction between the E1A full construct with the CBP-ID4 may shed light on the molecular basis of the interaction mechanisms between these two proteins. Therefore this appears an interesting interaction to be investigated to contribute to the understanding of how viral protein such as E1A and eventually E7 do interact with transcription factors.

3. The Methods

The study of the function of biomolecules is a highly interdisciplinary field. IDPs require specific skills to overcome the traditional way to analyze structured proteins (Uversky, 2011). In this section the different methods used in the present work are summarized, from genome browsing to samples preparation and analysis.

3.1. Bioinformatics analysis

GENOME BROWSING AND BIOINFORMATICS PREDICTION

The era of genome-scale biology made available an incredible amount of biological data with the concomitant proliferation of biological databases. Here we describe many tools used in this work for our research to understand the protein function. Proteins are the primary cellular effectors responsible for mediating the vast majority of functions/activities needed for the cell to function. Structural biology requires a large number of steps to convert DNA sequence information into protein including the selection of the proper expression constructs, the setting of the right growth conditions and efficient purification strategy. To reach this goal, several databases and web software tools are available for the scientific community. These are data banks containing information on DNA and protein sequences, expression profiles and protein ensembles and structures. A set of different tools used is summarized in table 3.1. The different results obtained with these tools for each protein studied in this work are reported in the following chapters.

Table 3.1. Bioinformatics tools used in this work and short description of their functionalities.

QQuery
Is a global query cross-database search engine that allows users to search many discrete health sciences databases at the National Center for Biotechnology Information (NCBI) website. [http://www.ncbi.nlm.nih.gov/gquery]

ViralZone
A web resource provides genomic and proteomic sequences as well as factsheets on all known virus families/genera (Hulo *et al*., 2011).

BLAST
Is the Basic Local Alignment Search Tool, designed to explore all the available protein and DNA sequence databases. It allows to find similar sequences from different genomes.

Clustal Omega
It allows multiple sequence alignment for DNA as well as for protein sequences. Multiple sequence alignments tools are essential in most bioinformatics analyses that involve comparing homologous sequences. It calculates the best match for the input sequences allowing the analysis of identities, similarities and differences on the output data (Sievers *et al*., 2011). From the virology viewpoint, this can be applied to compare the same protein belonging to different genomes in the same viral family showing the CRs, as well between different proteins from different viral families allowing better understanding of evolutionary biology.

ProtParam
Is a tool to preliminary analyzing the protein target sequence, one of the fist tools useful to get information about pI of proteins, as well amino acid propensity and to calculate the molar extinction coefficient (that helps measuring protein concentrations with UV absorbance at 280 nm). Is located at ExPASy server (Gasteiger *et al*., 2005).

SIGNALP
Is a post-translational modification predictor for signal peptides, glycosylation, phosphorylation and cleavage sites (Emanuelsson *et al*., 2007).

DISPHOS
Is a predictor of phosphorylation sites (Iakoucheva *et al*., 2004) that uses the Phospho.ELM phosphorylation database (Dinkel *et al*., 2011).

SLiMFinder
Is a predictor to find potential shared motifs in unrelated proteins using a model of convergent evolution and assigning a significance value to each motif (Davey *et al*., 2010) (http://bioware.ucd.ie/slimfinder.html).

TMHMM
Is a topology predictor of transmembrane region and their orientation (Krogh *et al*., 2001).

PSIPRED
Is a secondary structure prediction method (Buchan *et al*., 2013).

PROSO
Is a predictor of protein solubility (Smialowski *et al*., 2007) similar to **SolPro** (Magnan *et al*., 2009) and the **Prediction of Protein Solubility** in *E. coli* (Diaz *et al*., 2010).

MetalDetector
Is a predictor of disulfide bridges and metal binding sites (Passerini *et al*., 2011) similar to **DIANNA** (Ferrè and Clote, 2006) and **EDBCP** (Cheng *et al*., 2006).

PREDICTORS FOR INTRINSICALLY DISORDERED PROTEINS

More than 50 disorder predictors have been developed so far, since the first was created in Dunker's lab in 1996 (Romero *et al*., 1997). The properties of these predictors have been summarized and reviewed in several papers (Dosztányi and Tompa, 2008; Ferron *et al*., 2006; Uversky and Dunker, 2010). The majority of these predictors are available through public servers, and links to many of them can be found in the "Disordered Protein Database" (Disprot, http://www.dabi.temple.edu/disprot/index.php) (Sickmeier *et al*., 2007) the "Database of Disordered Protein Prediction" (D^2P^2, http://d2p2.pro) (Oates *et al*., 2013) and in the recently created IDPbyNMR website (http://www.idpbynmr.eu/home/). The intrinsic structural disorder is a widespread phenomenon, especially in eukaryotes, where conservative bioinformatics predictions suggest that 5%–15% of proteins are IDPs, and about 35-50% of proteins have IDRs longer than 30 residues (Ward *et al*., 2004a). It has been accepted that disorder is needed for signaling among the various living systems, increasing with the organism complexity (Dunker and Obradovic, 2001). Indeed, in mammals 75% of signaling proteins are predicted to contain long disordered regions (Dunker *et al*., 2008a).

A comparison between the amino acid sequences of IDPs and those of structured globular proteins showed that there are a number of significant differences, including amino acid composition, sequence complexity, hydrophobicity, aromaticity, charge, flexibility (Dunker *et al*., 2001). For example, IDPs are significantly depleted in hydrophobic (Ile, Leu, and Val) and aromatic (Trp, Tyr, and Phe) amino acid residues, which form and stabilize the hydrophobic core of folded globular proteins. These residues are called order-promoting amino acids. On the other hand, IDPs/IDRs are substantially enriched in polar (Arg, Gln, Glu, Lys, and Ser) and structure-breaking (Gly and Pro) disorder-promoting amino acid residues (Dunker *et al*., 2001; Radivojac *et al*., 2007). Among the 20 common amino acid residues, proline is the most disordered-promoting (Theillet *et al*., 2013). The predictors are based on different principles and can be classified into three main categories (Tompa and Fersht, 2009).

(*i*) **propensity-based predictors:** based on simple relative propensity of amino

acid residues. They use the difference in the probability for a given amino acid to be in regular secondary structure or to be in random coil. These are the cases of predictors "Globpot" (Linding *et al*., 2003b) and "Foldindex" (Prilusky *et al*., 2005).

(*ii*) **machine learning algorithms:** these are methods based and trained on datasets of disorder and order and evaluate intrinsic disorder on a per-residue basis. Those datasets are created from sequence and structural information existent on PDB files and their missing information from unfolded regions. They can use artificial neural networks (NN) like DisEMBL, good for short disordered regions (Linding *et al*., 2003a) or support vector networks (SVMs) like DISOPRED2 (Ward *et al*., 2004b) where the input data are generated by sequence alignment using PSI- BLAST, which is trained on a database of amino acids missing from PDB structures. PONDR® is a family of algorithms where both NN and SVMs types of machine learning algorithms are used (He *et al*., 2009).

(*iii*) **algorithms based on inter-residue contacts:** these are based on the inter-residue contacts essential for the maintenance of the folded protein structure. A level is attributed to those inter-residue contacts, thus the disorder of proteins is originated from the lack or low level of the inter-residue contacts, which cannot compensate the large decrease in conformational entropy during folding (Tompa and Fersht, 2009). This is the case of predictor IUPred (Dosztányi *et al*., 2005).

Thus it is now possible to predict the tendency of a polypeptide chain to be disordered based on its primary sequence with an approximate accuracy of 80% (He *et al*., 2009). The accuracy of predictors is regularly assessed as part of the critical assessment of structure prediction (CASP) experiment (Monastyrskyy *et al*., 2011). To get higher accuracies the best approach when analyzing a new protein sequence is to use metapredictors that combine the output of several individual predictors. This is the case of metaPrDOS (Ishida and Kinoshita, 2008), MeDor (Lieutaud *et al*., 2008), MD (Schlessinger *et al*., 2009) and PONDR-FIT (Xue *et al*., 2010). In this work, all the protein predictions presented were done using PONDR-FIT metapredictor.

MODELING

The development of comparative software allows to study the homology modeling of protein's three-dimensional structures (Eswar *et al*., 2006). With this method, the user provides a sequence alignment of the protein of interest and obtains a model based on known related PDB structures. For example by using MODELLER model containing all non-hydrogen atoms calculated by the comparative protein structure modeling based on satisfaction of spatial restrains can be obtained (Šali and Blundell, 1993). MODELLER also performs *de novo* modeling of loops in protein structures, multiple alignments of protein sequences and structures, optimization of various models of protein structure with respect to a defined objective, searching of sequence databases, clustering, comparison of protein structures etc.

3.2. Genes and Plasmids Generation

With the development of new instruments that facilitate the production of biological material, nowadays the synthesis of genes containing the DNA sequence of a target protein is feasible. Many companies like Invitrogen's GenArt®, OriGene, Eurofins MWG Operon, GenScrip, and DNA2.0 among others, provide web tools to order these genes, with several possibilities and strategies, including the optimization of the codons for the specific expression organism. For example, the strategy of over-expressing recombinant proteins in *E. coli* systems could be compromised resulting in low yields of expression if the open reading frame (ORF) of the protein contains infrequently codons used by *E. coli*, the so called "rare codons". In particular codons for arginine (AGG, AGA, CGA), isoleucine (ATA), leucine (CTA) and proline (CCC) should be avoided (Schenk *et al*., 1995). Nevertheless, the gene of interest can be also extracted from the organism of interest, if the corresponding cDNA is available. The cloning strategy should be designed carefully, as it could be the basis of a successful work. An incorrect/unlucky cloning strategy could compromise the behavior of the downstream process like protein solubility, yield, misfolding, etc.

The standard procedure to express a recombinant protein is to carry out a screen of different constructs to identify the most efficient for downstream purposes. The first step of the cloning process consists in the amplification of target gene from DNA template or plasmids through polymerase chain reaction (PCR) using specific primers. After purification, the amplified product is inserted into a specific expression vector. Different vectors may be selected in order to obtain native protein or protein fused to different tags. The tags are known in size and biological function and can enhance solubility or affinity for some specific purification systems or can allow the transport of protein in the periplasmic compartment (Esposito and Chatterjee, 2006). Later on those tags may be removed by proteolytic cleavage with specific enzyme, such as factor XA, enterokinase (EK) or (Tobacco Etch Virus) TEV protease (Arnau *et al*., 2006; Malhotra, 2009).

In the present work several plasmid were constructed or mutated to obtain the protein of interest. In particular the plasmid pET20/E7 containing the HPV 16 E7 gene was generously provided by Thomas Hey, Scil Protein GmbH, Halle, Germany (Mirecka *et al*., 2006); the E7 plasmids containing the silent mutations to improve the expression yield as well the plasmids of native E7 construct and the two E7 mutants for the cysteine residues C59A and C68T were prepared in this work using internal facilities. The plasmid pDEST17/E7CR3 containing the HPV16 E7 CR3 domain gene was prepared in this work using internal facilities.

The plasmid pET42/E1A13S containing the HAdV 2/5 hybrid 13S E1A gene was generously provided by Joe Mymryk, University of Western Ontario, London, Ontario, Canada (Pelka *et al*., 2011). The plasmid pETG-20A/E1A12S containing the HAdV 2 12S E1A gene was prepared in this work using internal facilities.

The plasmid pET200/CBPID4 containing the CBP ID4 domain was prepared in this work in close collaboration with Peter Tompa, VIB, Brussels. The plasmids used in this work are described in the next paragraphs, including the initial constructs and the protein sequences after proteolysis for tag removal.

PLASMID OF THE HPV 16 E7 FULL LENGTH - PET20/E7

The HPV 16 E7 construct is inserted in a pET20 (Novagen) plasmid that was prepared by classical restriction enzyme cloning. The HPV 16 E7 protein target is a native 98 residues. The final construct is not the native protein but a construct including a C- terminus 6xHis tag linked by 2 extra residues (LEHHHHHH). Thus, the final expression construct is a 106 residues protein. All vector polylinkers and map, as well the protein and DNA sequences are described on table 3.2.

Table 3.2. Datasheet of expression plasmid for HPV 16 E7. Description of vector map, target protein sequence and final expression protein and DNA sequences.

Protein sequence of interest: HPV 16 E7, 98 residues

MHGDTPTLHEYMLDLQPETTDLYCYEQLNDSSEEEDEIDGPAGQAEPDRAHYNIVTFCCKCD
STLRLCVQSTHVDIRTLEDLLMGTLGIVCPICSQKP

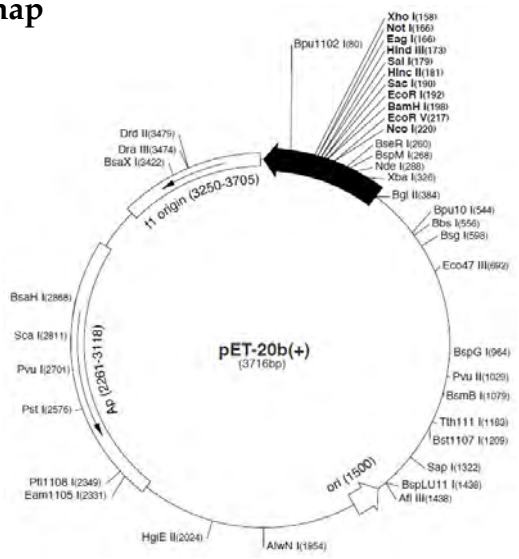
Protein sequence of expression plasmid HPV 16 E7, 106 residues

MHGDTPTLHEYMLDLQPETTDLYCYEQLNDSSEEEDEIDGPAGQAEPDRAHYNIVTFCCKCD
STLRLCVQSTHVDIRTLEDLLMGTLGIVCPICSQKP**LEHHHHHH**

DNA sequence (optimized)

ATGCA**C**GGAGATACACC**A**ACATTGCATGAATATATGTTAGATTTGCAACCAGAGACAACTGA
TCTCTACTGTTATGAGCAATTAAATGACAGCTCAGAGGAGGAAGATGAAATAGATGGTCCAG
CTGGACAAGCAGAACCGGACAGAGCCCATTACAATATTGTAACCTTTTGTTGCAAGTGTGAC
TCTACGCTTCGGTTGTGCGTACAAAGCACACACGTAGACATTCGTACTTTGGAAGACCTGTT
AATGGGCACACTAGGAATTGTGTGCCCCATCTGTTCTCAGAAACCA**CTCGAGCACCACCACC
ACCACCAC**

Novagen pET-20 map



PLASMID OF THE HPV 16 E7 CR3 DOMAIN - PDEST-17/E7CR3

The HPV 16 E7 CR3 domain construct is inserted in a pDEST-17 (Invitrogen) plasmid that was prepared by Gateway® cloning technology (Invitrogen). The HPV 16 E7 CR3 protein target is a native 54 residues. To this construct a N-terminus cassette was added to enable future tag removal abilities creating a pENTR construct, representing a protein with 77 residues. The construct including N-terminus TEV cassette was further inserted in a pDEST-17 plasmid, which confers the ability to have a 6xHis tag on the N-terminus, being the TEV recognition site for cleavage between the 6xHis tag and the target CR3 protein representing a protein of 98 residues. After the TEV proteolysis, the final construct will include 4 extra residues (GSFT) on N-terminus of CR3 representing a protein with 58 residues. All vector polylinkers and map, as well the protein and DNA sequences are described on table 3.3.

Table 3.3. Datasheet of expression plasmid for HPV 16 E7 CR3 domain. Description of vector map, target protein sequence and final expression protein and DNA sequences.

Protein sequence goal: HPV 16 E7 CR3, 54 residues
AEPDRAHYNIVTFCCKCDSTLRLCVQSTHVDIRTLEDLLMGTLGIVCPICSQKP

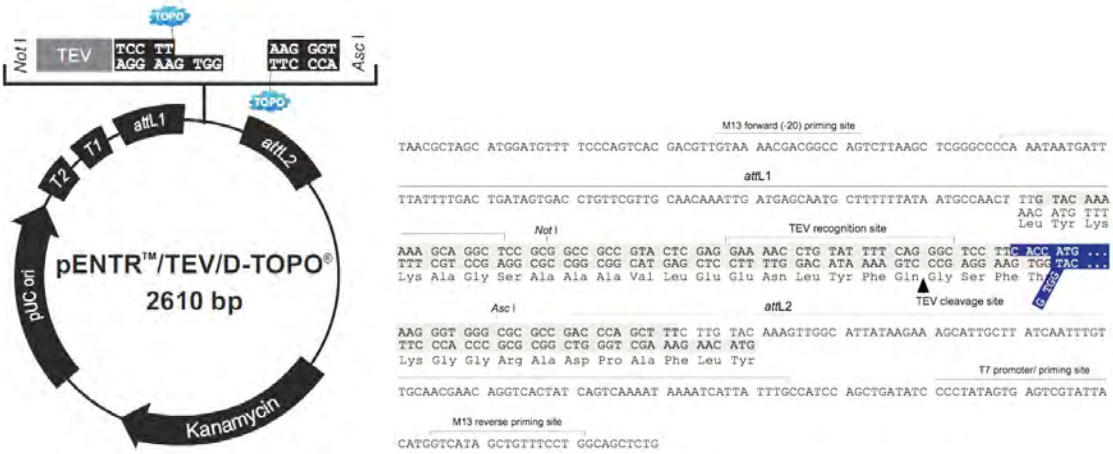
Protein sequence inserted on entry vector pENTR/TEV/D-TOPO, 77 residues
LYKKAGSAAAVL**EENLYFQG**SFTAEPDRAHYNIVTFCCKCDSTLRLCVQSTHVDIRTLEDLL
MGTLGIVCPICSQKP

Protein sequence of final expression vector including TEV cassette plus E7 CR3 transferred into pDEST-17, 98 residues
MSYY**HHHHHH**LESTSLYKKAGLYKKAGSAAAVL**EENLYFQG**SFTAEPDRAHYNIVTFCCKCD
STLRLCVQSTHVDIRTLEDLLMGTLGIVCPICSQKP

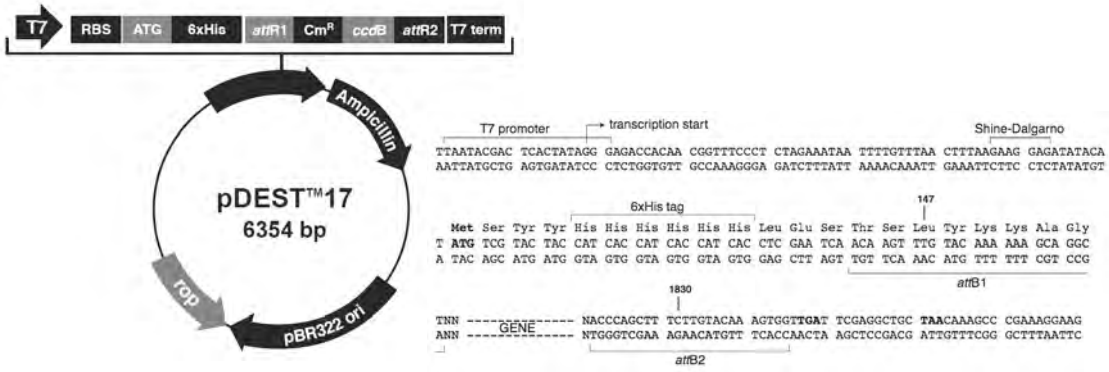
Protein sequence after TEV proteolysis: HPV 16 E7 CR3, 58 residues
GSFTAEPDRAHYNIVTFCCKCDSTLRLCVQSTHVDIRTLEDLLMGTLGIVCPICSQKP

DNA sequence of final construct after proteolysis
GGCTCCTTCACCGCAGAACCGGACAGAGCCCATTACAATATAGTAACCTTTTGTTGCAAGTG
TGACTCTACGCTTCGGTTGTGCGTACAAAGCACACACGTAGACATTCGTACTTTGGAAGACC
TGTTAATGGGCACACTAGGAATTGTGTGCCCCATCTGTTCTCAGAAACCA

Invitrogen pENTR/TEV/D-TOPO map and polylinker



Invitrogen pDEST-17 map and polylinker



PLASMID OF THE HADV 2/5 13S E1A - PET42/E1A13S

The HAdV 2/5 E1A13S construct is inserted in a pET42 (Novagen) plasmid that was prepared by classical restriction enzyme cloning. The HAdV 2/5 E1A13S protein target is a native 289 residues. The final construct is not the native protein but a construct including a C- terminus 8xHis tag linked by 2 extra residues (LEHHHHHHHH). Thus, the final expression E1A13S construct is a 299 residues protein. All vector polylinkers and map, as well the protein and DNA sequences are described on table 3.4.

Table 3.4. Datasheet of expression plasmid for HAdV 2/5 E1A. Description of vector map, target protein sequence and final expression protein and DNA sequences.

Protein sequence of interest: HAdV 2 E1A 13S 289 residues

MRHIICHGGVITEEMAASLLDQLIEEVLADNLPPPSHFEPPTLHELYDLDVTAPEDPNEEAV
SQIFPESVMLAVQEGIDLFTFPPAPGSPEPPHLSRQPEQPEQRALGPVSMPNLVPEVIDLTC
HEAGFPPSDDEDEEGEEFVLDYVEHPGHGCRSCHYHRRNTGDPDIMCSLCYMRTCGMFVYSP
VSEPEPEPEPEPEPARPTRRPKLVPAILRRPTSPVSRECNSSTDSCDSGPSNTPPEIHPVVP
LCPIKPVAVRVGGRRQAVECIEDLLNESGQPLDLSCKRPRP

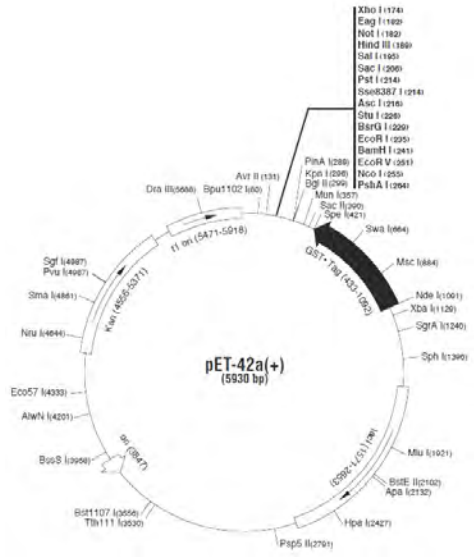
Final construct of E1A contains 299 residues - HAdV 2/5 hybrid

MRHIICHGGVITEEMAASLLDQLIEEVLADNLPPPSHFEPPTLHELYDLDVTAPEDPNEEAV
SQIFPESVMLAVQEGIDLFTFPPAPGSPEPPHLSRQPEQPEQRALGPVSMPNLVPEVIDLTC
HEAGFPPSDDEDEEGEEFVLDYVGHPGHGCRSCHYHRRNTGDPDIMCSLCYMRTCGMFVYSP
VSEPEPEPEPEPEPARPTRRPKLVPAILRRPTSPVSRECNSSTDSCDSGPSNTPPEIHPVVP
LCPIKPVAVRVGGRRQAVECIEDLLNEPGQPLDLSCKRPRP**LEHHHHHHHHH**

DNA sequence

ATGAGACATATTATCTGCCACGGAGGTGTTATTACCGAAGAAATGGCCGCCAGTCTTTTGGA
CCAGCTGATCGAAGAGGTACTGGCTGATAATCTTCCACCTCCTAGCCATTTTGAACCACCTA
CCCTTCACGAACTGTATGATTTAGACGTGACGGCCCCCGAAGATCCCAACGAGGAGGCGGTT
TCGCAGATTTTTCCCGAGTCTGTAATGTTGGCGGTGCAGGAAGGGATTGACTTATTCACTTT
TCCGCCGGCGCCCGGTTCTCCGGAGCCGCCTCACCTTTCCCGGCAGCCCGAGCAGCCGGAGC
AGAGAGCCTTGGGTCCGGTTTCTATGCCAAACCTTGTGCCGGAGGTGATCGATCTTACCTGC
CACGAGGCTGGCTTTCCACCCAGTGACGACGAGGATGAAGAGGGTGAGGAGTTTGTGTTAGA
TTAT**GTGGGGCACCCCGGGCACGGTTGCAGGTCTTGTCATTATCACCGGAGGAATACGGGGG
ACCCAGATATTATGTGTTCGCTTTGCTATATGAGGACCTGTGGCATGTTTGTCTACAGTCCT
GTGTCT**GAACCTGAGCCTGAGCCCGAGCCAGAACCGGAGCCTGCAAGACCTACCCGGCGTCC
TAAATTGGTGCCTGCTATCCTGAGACGCCCGACATCACCTGTGTCTAGAGAATGCAATAGTA
GTACGGATAGCTGTGACTCCGGTCCTTCTAACACACCTCCTGAGATACACCCGGTGGTCCCG
CTGTGCCCCATTAAACCAGTTGCCGTGAGAGTTGGTGGGCGTCGCCAGGCTGTGGAATGTAT
CGAGGACTTGCTTAACGAGCCTGGGCAACCTTTGGACTTGAGCTGTAAACGCCCCAGGCCA**C
TCGAGCACCACCACCACCACCACCAC**

Novagen pET-42 map



PLASMID OF THE HADV 2 12S E1A - PETG20A/E1A12S

The HAdV 2 E1A12S construct is inserted in a pETG-20A EMBL plasmid (Alzari *et al.*, 2006) that was prepared by Gateway® cloning technology (Invitrogen). The HAdV 2 E1A12S protein is a native 243 residues that was synthesized inside a pENTR vector (ordered from GeneArt® Invitrogen) containing the TEV recognition sequence on N-terminus and the kanamycin resistance sequence for further isolation of the target destination plasmids. This segment was then subcloned into the pETG-20A through the Gateway LR reaction, yielding a plasmid expressing the HAdV 2 E1A12S protein fused with thioredoxin A and a His tag at the N terminus. Using this approach the final construct after TEV cleavage yields the native 243 residues E1A12S protein. The E1A12S protein and DNA sequences are described on table 3.5.

Table 3.5. Datasheet of expression plasmid for E1A12S. Description of and DNA sequences.

Protein sequence of interest after TEV cleavage: E1A12S, 243 residues

MRHIICHGGVITEEMAASLLDQLIEEVLADNLPPPSHFEPPTLHELYDLDVTAPEDPNEEAV
SQIFPESVMLAVQEGIDLFTFPPAPGSPEPPHLSRQPEQPEQRALGPVSMPNLVPEVIDLTC
HEAGFPPSDDEDEEGPVSEPEPEPEPEPEPARPTRRPKLVPAILRRPTSPVSRECNSSTDSC
DSGPSNTPPEIHPVVPLCPIKPVAVRVGGRRQAVECIEDLLNESGQPLDLSCKRPRP

DNA sequence (native)

ATGAGGCACATCATCTGCCACGGCGGCGTGATCACCGAGGAGATGGCCGCCAGCCTGCTGGA
CCAGCTGATCGAGGAGGTGCTGGCCGACAACCTGCCCCCCCCCCAGCCACTTCGAGCCCCCCA
CCCTGCACGAGCTGTACGACCTGGACGTGACCGCCCCCGAGGACCCCAACGAGGAGGCCGTG
AGCCAGATCTTCCCCGAGAGCGTGATGCTGGCCGTGCAGGAGGGCATCGACCTGTTCACCTT
CCCCCCCGCCCCCGGCAGCCCCGAGCCCCCCCACCTGAGCAGGCAGCCCGAGCAGCCCGAGC
AGAGGGCCCTGGGCCCCGTGAGCATGCCCAACCTGGTGCCCGAGGTGATCGACCTGACCTGC
CACGAGGCCGGCTTCCCCCCCCAGCGACGACGAGGACGAGGAGGGCCCCGTGAGCGAGCCCGA
GCCCGAGCCCGAGCCCGAGCCCGAGCCCGCCAGGCCCACCAGGAGGCCCAAGCTGGTGCCCG
CCATCCTGAGGAGGCCCACCAGCCCCGTGAGCAGGGAGTGCAACAGCAGCACCGACAGCTGC
GACAGCGGCCCCAGCAACACCCCCCCCGAGATCCACCCCGTGGTGCCCCTGTGCCCCATCAA
GCCCGTGGCCGTGAGGGTGGGCGGCAGGAGGCAGGCCGTGGAGTGCATCGAGGACCTGCTGA
ACGAGAGCGGCCAGCCCCTGGACCTGAGCTGCAAGAGGCCCAGGCCC

DNA sequence (optimized for *E. coli*)

ATGCGCCATATCATTTGTCATGGTGGTGTTATTACCGAAGAAATGGCAGCAAGCCTGCTGGA
TCAACTGATTGAAGAAGTTCTGGCCGATAATCTGCCTCCGCCTAGCCATTTTGAACCGCCTA
CCCTGCATGAACTGTATGATCTGGATGTTACCGCACCGGAAGATCCGAATGAAGAAGCAGTT
AGCCAGATTTTTCCGGAATCAGTTATGCTGGCAGTTCAAGAAGGTATTGACCTGTTTACCTT
TCCGCCTGCACCGGGTAGTCCGGAACCGCCTCATCTGAGCCGTCAGCCGGAACAGCCTGAAC
AGCGTGCACTGGGTCCGGTTAGCATGCCGAATCTGGTTCCGGAAGTTATTGATCTGACCTGT
CATGAAGCAGGTTTTCCGCCTAGTGATGATGAAGATGAAGAAGGTCCGGTTTCAGAACCGGA
ACCTGAGCCTGAACCAGAGCCGGAACCGGCACGTCCGACCCGTCGTCCTAAACTGGTTCCTG
CAATTCTGCGTCGTCCGACCTCACCGGTTAGCCGTGAATGTAATAGCAGCACCGATAGCTGT
GATAGCGGTCCGAGCAATACCCCTCCGGAAATTCATCCGGTTGTTCCGCTGTGTCCGATTAA
ACCGGTTGCAGTTCGTGTTGGTGGTCGTCGTCAGGCAGTTGAATGTATTGAAGATCTGCTGA
ATGAAAGCGGTCAGCCGCTGGATCTGAGCTGTAAACGTCCGCGTCCG

PLASMID OF THE CBP-ID4 DOMAIN - PET200/CBPID4

The CBP-ID4 domain construct is inserted in a pET200 (Invitrogen) plasmid that was prepared by TOPO® cloning technology (Invitrogen). The CBP-ID4 protein target is a native 206 residues. This construct was added an N-terminus cassette that contains a 6xHis tag followed by an enterokinase (EK) recognition site for further tag cleavage. This construct represents expresses a protein with 242 residues. After tag proteolysis by EK, the final construct includes 5 extra residues (DHPFT) on the N-terminus of CBP-ID4 domain, representing a protein with 211 residues. All vector polylinkers and maps, as well the protein and DNA sequences are described on table 3.6.

Table 3.6. Datasheet of expression plasmid for CBP-ID4. Description of vector map, target protein sequence and final expression protein and DNA sequences.

Protein sequence of interest: CBP-ID4, 206 residues

QQQIQHRLQQAQLMRRRMATMNTRNVPQQSLPSPTSAPPGTPTQQPSTPQTPQPPAQPQPSP
VSMSPAGFPSVARTQPPTTVSTGKPTSQVPAPPPPAQPPPAAVEAARQIEREAQQQQHLYRV
NINNSMPPGRTGMGTPGSQMAPVSLNVPRPNQVSGPVMPSMPPGQWQQAPLPQQQPMPGLPR
PVISMQAQAAVAGPRMPSVQ

Protein sequence on final expression vector construct, 242 residues

MRGSHHHHHHGMASMTGGQQMGRDLYDDDDL**DHPFT**QQQIQHRLQQAQLMRRRMATMNTRNV
PQQSLPSPTSAPPGTPTQQPSTPQTPQPPAQPQPSPVSMSPAGFPSVARTQPPTTVSTGKPT
SQVPAPPPPAQPPPAAVEAARQIEREAQQQQHLYRVNINNSMPPGRTGMGTPGSQMAPVSLN
VPRPNQVSGPVMPSMPPGQWQQAPLPQQQPMPGLPRPVISMQAQAAVAGPRMPSVQ

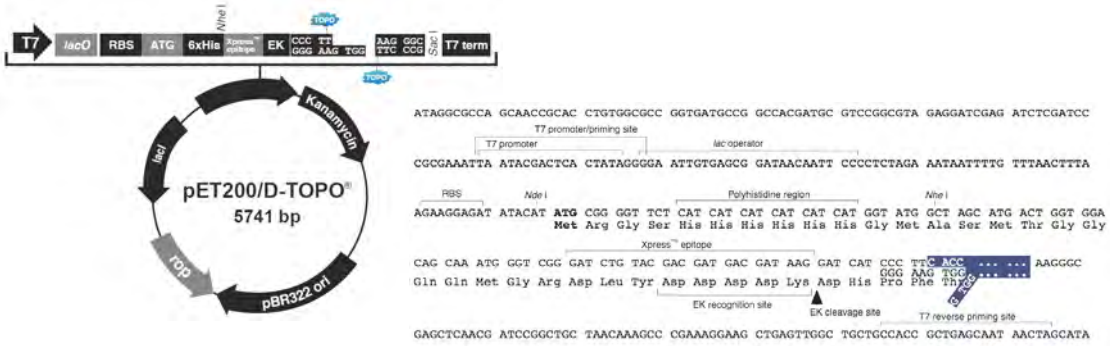
Protein sequence after proteolysis with EK protease, 211 residues

DHPFTQQQIQHRLQQAQLMRRRMATMNTRNVPQQSLPSPTSAPPGTPTQQPSTPQTPQPPAQ
PQPSPVSMSPAGFPSVARTQPPTTVSTGKPTSQVPAPPPPAQPPPAAVEAARQIEREAQQQQ
HLYRVNINNSMPPGRTGMGTPGSQMAPVSLNVPRPNQVSGPVMPSMPPGQWQQAPLPQQQPM
PGLPRPVISMQAQAAVAGPRMPSVQ

DNA sequence

GATCATCCCTTCACCCAGCAGCAGATCCAGCACCGCCTGCAGCAGGCCCAGCTCATGCGCCG
GCGGATGGCCACCATGAACACCCGCAACGTGCCTCAGCAGAGTCTGCCTTCTCCTACCTCAG
CACCGCCCGGGACCCCCACACAGCAGCCCAGCACACCCCAGACGCCGCAGCCCCCTGCCCAG
CCCCAACCCTCACCCGTGAGCATGTCACCAGCTGGCTTCCCCAGCGTGGCCCGGACTCAGCC
CCCCACCACGGTGTCCACAGGGAAGCCTACCAGCCAGGTGCCGGCCCCCCCACCCCCGGCCC
AGCCCCCTCCTGCAGCGGTGGAAGCGGCTCGGCAGATCGAGCGTGAGGCCCAGCAGCAGCAG
CACCTGTACCGGGTGAACATCAACAACAGCATGCCCCCAGGACGCACGGGCATGGGGACCCC
GGGGAGCCAGATGGCCCCCGTGAGCCTGAATGTGCCCCGACCCAACCAGGTGAGCGGGCCCG
TCATGCCCAGCATGCCTCCCGGGCAGTGGCAGCAGGCGCCCCTTCCCCAGCAGCAGCCCATG
CCAGGCTTGCCCAGGCCTGTGATATCCATGCAGGCCCAGGCGGCCGTGGCTGGGCCCCGGAT
GCCCAGCGTGCAG

Invitrogen pET-200 map and polylinker



CLONING STRATEGIES

The classic method to insert an amplified PCR product into a vector is to use restriction enzymes that cleave DNA at specific recognition sites. Both DNA and cloning vector have to be treated with two restriction enzymes that create compatible ends. Later on those ends are connected by a ligation reaction performed by the bacteriophage T4 DNA ligase. Finally an aliquot of the product of the reaction is transformed in DH5α *E. coli* competent cells and positive clones are screened both by PCR in DNA sequencing. However the classic cloning strategy sometimes is not feasible to prepare different constructs in parallel because not all host vectors and target genes can be cleaved by the same restriction enzymes. Thus, other cloning strategies are available exploiting ligation-independent cloning, such as Gateway® (Invitrogen) and the most recent Electra™ (DNA2.0) (Katzen, 2007). Gateway® cloning technology has been one of the most used strategies allowing simultaneous preparation of plasmids for different expressions systems strategies. Basically is consists in the generation of an entry clone that can be recombined in several compatible vectors without the use of any restriction enzyme. There are many ways to create an entry clone but the most straightforward method is by adding four bases to the forward primer (CACC) to create blunt-end PCR products for directional TOPO® cloning. The ligation reaction of the PCR product to the pENTR vector is performed by the topoisomerase I. After isolation of the entry clone, the second step is to generate an expression vector. This is done by an exchange between the gene

on the entry clone with the final expression vector, performed by LR Clonase®. Because each vector contains different antibiotics resistance, the isolation of the final expression vector can be easily obtained by transformation of E. coli competent cells and selection on specific resistance plates (Figure 3.1). A large selection of Gateway expression vectors is available, with the possibility to express the native target protein as well as conferring the possibility to include tags on its N-terminus or C-terminus. Those tags can be only histidine tails for helping the purification strategies, or even well known fusion proteins that confer solubility and can help on increasing expression yield.

Figure 3.1. Gateway cloning system. The exchange of genes between entry and destination vectors is performed by the LR® Clonase®. Each vector contains different resistance genes to allow selective isolation of the final expression vector.

In the present thesis, the cloning system used to create the entry clone for HPV 16 E7 CR3 domain was the Gateway Cloning System pENTR/TEV/D-TOPO. This version of Gateway Cloning System includes the existence of a cassette of TEV recognition site on the N-terminus of the protein. Thus, any destination expression vector containing tags on the N-terminus may be removed by proteolysis using TEV protease. The following expression vectors have been used for the creation of simultaneous expression vectors: pDEST-17 (conferring 6x histidine N-terminus), pETG-30A (conferring GST plus 6x histidine N-terminus), and pDEST-His-MBP (conferring MBP plus 6x histidine N-terminus). It is important to highlight that using this method, the final purified protein will include extra residues on the N-terminus. For example using the expression vector pDEST-17 that contains the 6xHis tag, the expressed and purified target protein will have extra 44 amino acids. After tag removal using TEV protease, the final construct will contain 4 extra residues on the N-terminus, GSFT.

For the creation of the expression vector for CBP-ID4 protein, a similar strategy was used. This work was done on Peter Tompa's lab where the entire

gene of the big 2442 amino acid CBP protein is available. To easily access the different domains present in CBP (Figure 2.4), a library of primers to read the specific region one wants to clone is available. Those primers were created thinking on the TOPO Cloning System strategy and therefore every forward primer contains the CACC sequence at the 5' end of the primer, to enable directional cloning. Due to experience with similar domains of CBP protein, the blunt end DNA constructs created by PCR are directly ligated into a final expression vector pET200/D-TOPO avoiding the generation of entry clones that need afterwards to be exchanged into destination expression vectors. This methodology as vantages and disadvantages because from one side it is a really fast method to get an expression plasmid, but on the other hand there are no ways to try other vectors containing different tags. The pET200/D-TOPO confers 6xHis tag on the N-terminus that may be cleaved due to the presence of enterokinase (EK) recognition site. It is important to notice that using this methodology the final construct before cleavage will contain 36 extra amino acids an the N-terminus. After removing the tag by using EK protease, the final construct will keep 5 extra residues on the N-terminus, DHPFT.

MUTAGENESIS

Site-directed mutagenesis is a standard technique to improve the final expression vector, or to study different point mutations as a strategy to understand downstream protein structural and functional modifications. This method is used to make point mutations, replace amino acids, and delete or insert single or multiple adjacent amino acids. The point mutations that exchange a single nucleotide, but the new codon specifies the same amino acid are called silent mutations. Basically it codes the same amino acid, being an easy way to avoid "rare codons" that may decrease the expression yield. In the present work all mutagenesis were performed using the QuikChange II Site-directed Mutagenesis kit, Agilent Technologies. This method uses basically the double-stranded DNA vector template containing the target gene and two complementary synthetic oligonucleotide primers, both

containing the desired mutation. The primers are mixed with the DNA vector template and extended during PCR cycles performed by a high fidelity DNA polymerase (PfuUltra, Stratagene). The PCR product is then treated with DpnI (Figure 3.2), an endonuclease that will digest the DNA template due to its high specificity to Dam methylated and hemimethylated DNA isolated from *E. coli* strains. The new copies of DNA PCR product were never methylated and thus are not digested. The digested solution is then transformed into XL1-Blue super-competent cells and subsequently subjected to sequencing analysis.

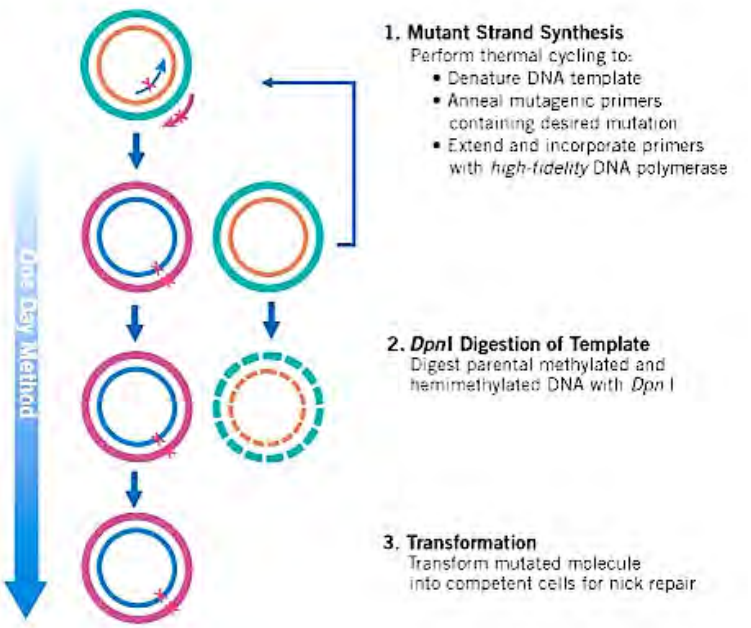


Figure 3.2. Overview of the QuikChange II Site-directed Mutagenesis Method adapted from Agilent Technologies.

The expression vector pET20-E7 containing the HPV 16 E7 gene was generously provided by Thomas Hey, Scil Protein GmbH, Halle, Germany (Mirecka *et al.*, 2006). The sequencing analysis of this vector showed the sequence corresponding to the native DNA of E7 protein. Using the Biomax ProteoExpert tool, it was found that the native codon for His-2 might allow the possibility of alternative translation initiation and premature termination of translation. Moreover the native codon for Pro-6 might induce low expression yields in *E. coli*.(Mirecka *et al.*, 2006). Two silent mutations in a

single site-directed mutagenesis reaction were preformed to avoid those two situations. The His-2 was mutated from CAT to CAC and the Pro-6 was mutated from CCT to CCA. To generate the optimized plasmid containing the E7 gene PCR was performed using the forward primer 5′-GGAGATATACATATG**CAC**GGAGATACA**CCA**ACATTGCATGAATATAT G-3′ and the reverse primer 5′-CATATATTCATGCAATGT**TGG**TGTAT CTCC**GTG**CATATGTATATCTCC-3′ designed to generate both silent mutations for His-2 and Pro-6 in a single PCR reaction. The optimized plasmid was transformed into *E. coli* XL1-Blue (Stratagene) competent cells and the resulting clones were screened by DNA sequence analysis. The optimized plasmid was replicated and used for all the HPV 16 E7 full-length protein samples preparations presented in this work.

Site directed mutagenesis was used as well to create 3 different constructs of the full-length HPV 16 E7 protein: the mutants C59A and C68T, and the native HPV 16 E7 protein without any tags. The E7 mutant C59A was generated using the forward primer 5'-CAATATTGTAACCTTTTGT**GCC**AAGTGTG ACTCTACGCTTC-3' and the reverse primer 5'- GAAGCGTAGAGTCACA CTT**GGC**ACAAAAGGTTACAATATTG-3'. The E7 mutant C68T was done using the forward primer 5'- GACTCTACGCTTCGGTTG**ACG**GTAC AAAGCACACACG-3' and the reverse primer 5'- CGTGTGTGCTTTGTA C**CGT**CAACCGAAGCGTAGAGTC-3'. The native HPV 16 E7 protein was generated by the introduction of a stop codon TAG right after the codon for the E7 protein C-terminus residue, Pro98, thus removing the C-terminus 6xHis tag LEHHHHHH. For this purpose the forward primer 5'-GTTCTCAGAAACCA**TAG**CTCGAGCACCACCACC-3' and the reverse primer 5'-GGTGGTGGTGCTCGAG**CTA**TGGTTTCTGAGAAC-3' were used (regions of codon mutation/insertion are highlighted).

3.3. Sample Preparation

This section describes all laboratory practices done since the use of generated plasmids till the preparation of protein samples necessary for further characterization.

RECOMBINANT PROTEIN EXPRESSION IN *E. COLI.*

Currently the number of recombinant proteins used for research applications has increased dramatically. Several host systems are available for protein production, including fungi, plants, bacteria, insects, yeast and mammalian cells (Shatzman, 1995). The selection of an expression system for high-level production of recombinant proteins depends on many factors, including biological activity of the target protein, post-translational modifications, cells growth features, intracellular and extracellular characteristics and expression levels. Production of these proteins has a remarkable demand on structural biology research and is crucial to the present work. *Escherichia coli* (*E. coli*) offers a means for the rapid and economical production of recombinant proteins (Tong *et al*., 2008). A wealth of biochemical and genetic knowledge of *E. coli*, have driven the development of a variety of strategies for achieving high-level expression of protein. The major challenges for obtaining the high yield of protein at low cost involves several aspects such as expression vectors design, gene dosage, transcriptional regulation (promoter), mRNA stability, translational regulation (initiation and termination), host design considerations, codon usage, and fermentation factors available for manipulating the expression conditions (Jana and Deb, 2005). Different *E. coli* strains, each one with specific features for different expression systems, were used in the present work.

Once selected the expression system, the strategy to understand the optimal expression conditions needs to be tested and is subjected to several variations. Thus, the best approach is to use parallel test expression strategies (Lesley, 2009). A preliminary expression test can be performed using small volume scale to find the best conditions to be reproduced in large volumes in order to get soluble recombinant protein (Hunt, 2005). The purification trials should be well designed taking in consideration all the typical variant conditions like, culture medium, temperature, optical density, isopropyl β-D-1-thiogalactopyranoside (IPTG) inducer concentrations, end expression times. A library of different *E. coli* strain is available each one with specific advantages to overcome typical difficulties. The *E. coli* strains used in the present work were: BL21(DE3); BL21(DE3)pLysS, a protease deficient strain,

Rosetta(DE3) for genes containing rare codons; Origami(DE3) for proteins containing disulfide bridges and Gold(DE3) for increasing expression yields. Typically one can try one *E. coli* strain and try two different optical densities (0.6 and 1.0), two IPTG concentrations (0.1 and 1.0 mM), three different expression temperatures (17, 30 and 37 ºC), three expression times (4h, 8h and overnight). Cells should be harvested and disrupted in equal conditions and solubility of the target protein should be checked by SDS polyacrylamide gel (SDS-PAGE). In case of non-satisfactory results, one can try different strains of *E. coli* cells. If all the trials to get soluble protein fail, the insoluble fraction can be recovered performing refolding trials. The composition of the culture medium is one of the most important conditions to take in consideration and should be planned carefully. The additives to be added can influence the yield and solubility of the protein expression. For example, proteins coordinating metal ions like zinc-fingers need extra-addition of zinc (II) into the culture medium. The amount of carbon and nitrogen source can also influence the yield. Also changing the carbon and nitrogen sources to isotopic labeled ones may change the protein expression yield. Moreover, culture medium containing deuterated water may need cells adaptation due to its potential toxicity.

PROTEIN PURIFICATION

The protein purification strategy depends always on the proteins' biophysical and biochemical properties. Several characteristics must be considered whether the protein is soluble and, if not, which agents might help to solubilize it. Is the protein sensitive to high or low salt concentrations, high or low pH, high or low temperatures, or oxidation, particularly by oxygen? Is the protein labile at high or low concentrations? A set of preliminary experimental trials should be done to learn which reagents might be present during the purification, which must be avoided, and under which conditions the protein has to be stored. Designing the protein purification strategy for a new protein starting from scratch requires preliminary bioinformatics study to predict some biochemical and biophysical characteristics and also to compare with homologous proteins purification strategies published in the

literature, helping to draft a purification plan (Hunt, 2005). For example, if the protein contains several free cysteine residues, anaerobic purification methods and the use of reducing agent may be required to avoid the formation of non-physiological multimers and oligomers. Besides each protein has particular characteristics, the first protein purification step will depend mainly upon the localization of the produced protein within the host, considering if it was produced within the cytoplasm either as soluble or insoluble protein, the so called inclusion bodies (IBs), transported into the periplasmic space, enclosed into the cell membranes or even excreted into the growth medium (Linn, 2009). For each case the isolation is performed in different ways. Assuming the protein was not excreted into the growth medium, the cell lysis is the first step for protein purification. The disruption of the cells can be performed though several techniques, spanning from physical to detergent-based methods. For instance the French-press is an efficient method, though heating the final lysate, while freeze-thaw or enzymatic and detergent lysis are considered mild methods but less efficient. In the present work, the cell disruption was performed by sonication, which comprises pulsed, high frequency sound waves to mechanically agitate and lyse the cells. The isolation of protein using sonication should be done carefully because the mechanical waves energy will heat the sample. To avoid sample warming, the sonication should be carried using an ice-bath and performed with short pulses with intervals to allow for temperature decreasing. Sonication methods can be also performed inside an anaerobic chamber (glove-box) for anaerobic purification strategies.

Once the impure protein fraction is isolated, the following purification step can be performed in a multitude of chromatography runs. The methodology should be always optimized to reach an efficient protocol in terms of yield, speed and costs. Among all the different chromatography techniques, in this work will be described the three most used technique: immobilized metal ion affinity chromatography (IMAC), ion exchange chromatography (IEX) and size exclusion chromatography (SEC).

The immobilized metal ion affinity chromatography (IMAC) is currently the most used affinity technique exploiting the interaction between chelated transition metal ions (like Zn^{2+} or Ni^{2+}) and the side chains of specific amino

acid (like histidines) on the protein. Basically using IMAC the target protein is tightly bound to the resin matrix and the impurities washed out with increasing concentrations of imidazole, which acts like a competitive agent. At higher imidazole concentrations the target protein is also eluted, almost pure (Block *et al*., 2009). The tags may be cleaved afterwards depending on the present of protease specific recognition sites cloned in the expression vector designed. Enzymatic digestions can be performed by proteases like TEV, Enterokinase, Thrombin etc., according to the strategy chosen. Thus, a second IMAC reverse chromatography is performed in order to separate the cleaved target protein from the tags.

Ion exchange chromatography (IEX) separates proteins on the basis of a reversible interaction between the polypeptide chain and a specific charge ligand attached to a chromatographic matrix. The isoelectric point (pI) of the target protein must be known. The sample is loaded in conditions favoring specific binding like carefully calibrated pH and low ionic strength salt concentration, in order to enhance the interaction between target protein and column matrix. The unbound impurities are washed out and the bound protein is eluted changing the pH or the ionic strength of the elution buffer. If the overall net charge of the protein is positive, an cationic IEX resin must be used; thus if its negative, an anionic IEX resin must be used. The buffer pH should be at least ± 1 unit different from the protein pI.

Size exclusion chromatography (SEC) is a separation technique based on hydrodynamic radius of the proteins. The column matrix is composed by precision sized beads containing pores of given sizes. Larger proteins whose hydrodynamic dimensions are big to fit inside any pore will have access only to the mobile phase between the beads, being excluded as they just follow the solvent flow and reach the end of the column before molecules with smaller size. Proteins whose hydrodynamic dimensions are smaller will be drawn into the pores by diffusion, having access to the mobile phase inside and between the beads. Thus, smaller molecules had long ways to cross with several small retention times between the diffusion movements through the beads pores. Due to largest retention smaller hydrodynamic molecules will elute last during the size exclusion separation. SEC can be used to separate protein by size and shape, to exchange the buffer and also to isolate protein

mixtures, to separate monomers, multimers or oligomers. In case of folded proteins, it can be also used to determine the molecular mass performing a molecular weight distribution analysis using available standards.

Therefore SEC profiles are dependent on the hydrodynamic volume of a protein being one of the most important and fundamental structural parameters of a protein molecule. Hydrodynamic volume is a prerequisite for an accurate classification of a protein conformation. It changes dramatically depending if the protein hydrodynamic dimension is compact like a folded protein or extended or partially extended like an IDP.

Comparing two proteins with the same molecular weight, a well folded protein will have a smaller hydrodynamic radius while an IDP will have bigger hydrodynamic radius behaving like a big folded protein on SEC. Thus, the SEC retention times for a folded protein and an IDP of same molecular mass will be different and the IDP will elute first. SEC has been used early for separation of unfolded and folded proteins (Gupta, 1983). However due to the particular characteristics of IDPs, SEC can be used for analytical study the conformational IDP properties in solution where size and shape of molecules are the prime separation parameters (Uversky, 2013a). In the present work, SEC was performed as the last step purification for the preparation HPV 16 E7 and HAdV 2/5 13S E1A proteins. A HiLoad 16/60 Superdex 75 column with a sample loop of 5.0 ml was used connected to a water thermostatic cooling system at 4 ºC. The details of the SEC protein purification are reported in the results chapter of the present work.

REFOLDING

Purification and refolding a protein from the insoluble fraction could be a challenging task and should be planed carefully. The strategy can be based on a previous bioinformatics analysis of the target protein combined with the state of art similar protein systems. Although each protein is unique requiring specific treatments, the literature is full of successful protocols summarized in many reviews and book chapters (Burgess *et al*., 2009; Cowieson *et al*., 2006; Qoronfleh *et al*., 2007; Singh *et al*., 2005; Vincentelli *et al*., 2004).

The first step on refolding is to solubilize the IBs. The solubilizing agent/denaturant could be a chaotropic agent like GuHCl and urea, or a detergent like SDS and it should be prepared in controlled pH buffer. As for the other technique strategies already described in this work, the key concept for refolding is the systematic, parallel screening of multiple refolding conditions. Many additives may prove useful in refolding, but to prevent aggregation and precipitation upon refolding is crucial to refold at low protein concentration. However, many variables in refolding should be controlled such as pH, temperature, salt concentration, redox environment, and the presence of divalent metal ions.

Redox agents, various redox pairs can be used including reduced and oxidized cysteine or glutathione as well as reducing agent like β-mercaptoethanol (BME), dithiothreitol (DTT) or tris (2-carboxyethyl) phosphine hydrochloride (TCEP) to control the oxidation state of the protein. Since the cytoplasm of *E. coli* is very reducing, most internal proteins are in the reduced state. If the protein as native disulfide bonds, the molar ratio of reduced to oxidized form is sometimes varied to achieve optimal native disulfide bond formation keeping native free cysteine residues. However if the protein doesn't have native disulfides bonds, and higher NMR samples concentration are required, the high concentration allow better intermolecular interaction and the presence of free cysteine residues may induce incorrect disulfide bridges prone to aggregation states. In this cases higher amount of DTT and TCEP are required.

Salt concentration, on range from 50-100mM is used to increase solubility of the native protein. Often 50–100 mM salt is used, when removing the IBs solubilizing agent.

Chaotropic agents, to avoid aggregation during refolding, 1 M urea and 0.5 M GuHCl have been used in refolding solution to dissociate possible aggregates due to interactions among partially folded intermediates.

pH, in general the buffer pH should be at least one pH unit away from the pI to avoid a zero net charge of the protein becoming prone to precipitation. Some protocols relies only on single pH refolding procedure (Coutard *et al*., 2012).

Temperature, most refolding procedures are carried at room temperature,

which is low enough to prevent thermal damage to the protein and high enough to increase the thermal motion of the molecules important to reach its native state.

Proline and Arginine, proline is considered a osmoprotectant and has been found to be effective increasing solubility both in vitro and in vivo (Ignatova and Gierasch, 2006). Arginine can decrease aggregation by slowing the rate of protein–protein interactions by supramolecular assemblies formations in solution. However the effective concentrations are reported in a range of 0.5–1.0 M (Das *et al*., 2007).

Glycerol, has been found to be an excellent refolding additive in many cases, usually used in the 5–30% range.

Detergents, can prevent aggregation during refolding. At low concentrations they bind weakly to exposed hydrophobic regions preventing aggregation.

As their concentration decreases they dissociate and allow reformation of native structure. At high concentration detergents are denaturants, but at low concentration they can act as an artificial chaperone promoting refolding without aggregation. An important aspect about detergents is the Critical Micelle Concentration (CMC), the concentration at which micelles begin to form. The CMC value depends on the buffer used, on pH and on ionic strength. For each buffer system the CMC must be determined or obtained from tabulated values (Brito and Vaz, 1986; Jumpertz *et al*., 2011).

Once the protein of interest is solubilized and all the refolding buffer conditions have been defined, the refolding can be attempted. The refolding procedure is the removal of the denaturant agent allowing the protein to reach its native state. Thus, refolding can be performed by dilution, multi-step dialysis, single dialysis, or with on-column refolding. Dialysis is one of the most used methods, while the on-column refolding can exploit the presence of a tag on the protein for affinity, binding to the column, allowing the removal of denaturant agent by washing steps. If it doesn't precipitate inside the column, the refolded protein can be further eluted. On the other hand, the dilution refolding can be performed by: **Reverse dilution**, addition of refolding buffer to denatured protein with mixing between each addition; **Flash dilution**, addition of denatured protein to refolding buffer quickly) and **Drip dilution,** addition of denatured protein to refolding buffer very slowly,

drop-by-drop, allowing refolding at low concentration.
Several commercial products have been developed to help identifying suitable refolding conditions like EMD/Novagen's iFOLD kits, Pierce Biotechnology's ProMatrix™ and AthenaES's QuickFold™.
The HPV 16 E7 CR3 domain was tested in many different plasmids constructs including only 6xHis tag as well other fusion protein tags like GST and MBP ton increase E7 CR3 solubility. All the expression tests using those different constructs showed E7 CR3 domain to be expressed in insoluble phase as IBs. The construct containing only the N-terminus 6xHis tag was chosen for a scale-up protein expression and purification. The refolding trials were carried using all the trials described for here refolding. The results and description of the successful protocol for the refolding of HPV 16 E7 CR3 domain are described in the results chapter of the present work.

SAMPLE HANDLING

The preparation of samples for many different biophysical and biochemical characterization should be prepared according to technical limits of each technique.
Test expression, is a preliminary assay to learn the best conditions for a scale-up downstream process. To avoid different expression results, the medium where those tests are done should be identical to the scaled-up final medium. For instance, a test using LB rich medium may present different results from a test using M9 minimal medium, special due to the controlled additives one can choose for the M9 medium preparation (like the addition of a specific metal essential for protein correct folding as in the case of zinc-proteins). When isotopically labeling a protein, the sources of carbon and nitrogen are different from the ones used for test expression and in some cases may lead to different expression yields.
Zinc additive, zinc is essential for many cellular processes, including DNA synthesis, transcription, and translation, but excess can be toxic (Babich and Stotzky, 1978; Kindermann *et al*., 2005). To find the optimal amount of zinc additive, different trials should be performed using controlled minimal medium instead of rich LB (Outten and O'Halloran, 2001). The amount of zinc

was tested comparing cultures depleted of zinc to a concentration range of 10 - 400 μM of $ZnCl_2$ and $ZnSO_4$ additions.

Reducing agents, can be exogenous sulfhydryl containing reducing agents like DTT, or non-sulfhydryl reducing agents like TCEP. The optimal reducing activity of DTT is on 6.5 - 9.0 pH range, white TCEP as a wide optimal reducing activity pH range spanning from 1.5 - 9.0. In some cases was reported TCEP to be more useful than DTT for protein sample preparations (Getz *et al*., 1999; Krezel *et al*., 2003). In the present work the control of the redox state of the proteins was crucial. Three different reducing agents and their concentration were compared and analyzed.

Anaerobic purification, protein containing cysteine residues must be handled inside an anaerobic chamber under nitrogen atmosphere to prevent oxidation of cysteine residues. With exception for CBP-ID4 protein, all buffers used in this work were extensively degassed with nitrogen and argon and used for all purification steps and sample preparation. Reducing agents were added to keep the cysteine residues reduced.

Samples of different proteins

HPV 16 E7 protein was prepared as unlabeled, ^{15}N labeled and $^{13}C,^{15}N$ labeled, to accomplish its NMR characterization.

E7 cysteine mutants, were done using E7 6xHis tag sequence as template. The C59A and C68T mutants were expressed and purified in identical way as for E7 6His tag protein without any mutations. The proteins where compared by NMR and no significant changes were observed when comparing with the non-mutated samples.

E7 native protein was generated by the insertion of a stop codon right before the 6xHis tag. The protein was purified three runs of size exclusion chromatography purification. The buffers used were prepared in identical conditions as those used for 6xHis tag E7 sample. The comparison of pure protein samples was performed by NMR spectroscopy, and no significant changes were observed when compared with the E7 including the 6xHis tag.

Metal removal, was performed for HPV 16 E7 sample, by removing zinc to prepare E7 apoprotein using 10mM EDTA as chelating agent on degassed 10mM HEPES buffer pH 8.0, 50mM KCl, and 10 mM DTT reducing agents.

E7 CR3 protein was prepared to help in the assignment of the folded region due to the difficulties to observe this domain on the full-length E7 protein. It was not possible to overcome the concentration of 40μM of E7 CR3.

HAdV 2/5 E1A protein was prepared as unlabeled, ^{15}N labeled and ^{13}C,^{15}N labeled, to accomplish its NMR characterization.

CBP-ID4 has been expressed and purified and the first HSQC acquired. Different trials are still under investigation in order to increase the concentration yield and avoid observed sample degradation.

Samples for different techniques

MS samples, the mass spectrometry E7 and E1A proteins samples were prepared using of unlabeled carbon and nitrogen sources as for CD DLS and SAXS. Samples were purified using the same protocol to ensure the use of degassed buffers and the fresh addition of 10mM DTT to keep the cysteine residues reduced. For MS technique some specific buffers should be avoided so the final MS samples was dialyzed against degassed 10mM ammonium sulfate at pH 6.6 using sealed containers. It has been reported that cysteine residues can be oxidized to Cys-SOH which can be observed by MS (Conte and Carroll, 2013; Ratnayake *et al*., 2013; Rehder and Borges, 2010; Shetty *et al*., 2007).

CD samples, of HPV 16 E7 and HAdV 2/5 E1A proteins where prepare in anaerobic conditions. For CD technique some specific buffers should be avoided so the final CD samples were dialyzed against 5mM potassium chloride at pH 7.5 using sealed containers with addition of fresh 10 mM DTT.

DLS samples, of HPV 16 E7 protein were prepared in the same conditions as for the other biophysical techniques like SAXS and NMR. The final buffer was 10mM Hepes pH 7.5, 150mM KCl, 10mM DTT, 10μM $ZnCl_2$.

SAXS samples, of HPV 16 E7 protein were prepared and carried to the SAXS facility at EMBL X33 beam line storage ring DORIS III on DESY, Hamburg in close collaboration with Dmitri Svergun and colleagues. Sample were prepared in the same buffer used for NMR measurements, 10mM Hepes pH 7.5, 150mM KCl, 10μM $ZnCl_2$, 10mM DTT. Samples were prepared in a concentration range from 1.3 - 2.7 mg ml^{-1} (108 - 223 μM) was used.

Concentrated samples where prepared right before the measurements using a refrigerated centrifuge.

Samples for NMR

^{15}N labeled samples were prepared to screen different conditions by NMR such as pH, temperature and sample concentration. **pH titrations** experiments were done by dilution with desired pH buffer followed by concentration steps using Millipore centrifugal concentrator device at 4ºC. Additions of fresh DTT and Roche protease inhibitors were done right before their use for all concentration steps. The final sample pH was checked on the concentrator flow-through buffer. Temperature experiments were carried in a range from 4ºC to 40ºC. For **concentration dependence** the NMR samples were analyzed in a stepwise concentration screening, to avoid aggregates and oligomer formation and screen the optimal concentration value. **^{13}C,^{15}N** samples were prepared to NMR experiments using carbon detection to help the sequence assignment of the different protein in study. **^{2}H,^{13}C,^{15}N** samples were prepared to understand the oligomer formation as the case of E7 protein. The deuterated NMR samples were prepared starting with the adaptation of *E. coli* cells transformed with pET20/E7 in deuterated medium done stepwise from 20 - 90% deuterium. The final expression medium was Silantes® OD2 CON triple-labeled medium. The purification was performed in identical way as done for unlabeled, ^{15}N and ^{13}C,^{15}N samples for further NMR experiments comparison. **SedNMR samples**, were prepared using a double labeled HPV 16 E7 sample concentrated till 15 mM and loaded on a 3 mm rotor.

3.4. Biophysical and Biochemical Characterization

UV-VISIBLE SPECTROSCOPY

The ultraviolet (UV)–visible spectroscopy applied for protein research field can be used for many application including detection of eluting components in high performance liquid chromatography (HPLC), determination of the oxidation state of a metalloprotein center of a cofactor as well the determination of the maximum absorbance of proteins and DNA for measurement of their concentrations. The DNA quantification is based on the purine and pyrimidine absorption that has maximum around 260 nm. Proteins contain aromatic amino acid residues, which absorb light in the UV-range with absorption maxima around 280 nm. These amino acids are tryptophan (Trp), tyrosine (Tyr), and phenylalanine (Phe). The Trp, Tyr, and Phe absorption spectra of IDPs can be compromised as typically IDPs are depleted of these residues (Dunker *et al*., 2001).

The concentrations can be calculated using the Beer-Lambert law, $\varepsilon = \frac{A}{c.l}$, where A is the absorbance value at the chosen wavelength, c the sample concentration (M) and l length of the light path through the sample (cm). The absorption spectrum of a protein is usually normalized to a concentration unit and cell width, thus depending on protein specific molar absorption coefficient ε and the measured wavelength λ absorption values. The protein molar extinction coefficient can be calculated, for example using the ExPASy Protparam tool (Gasteiger *et al*., 2005) [http://web.expasy.org/ protparam/]. The theoretical molar extinction coefficients of the proteins studied in this work are reported on table 3.7, considering all cysteine residues are reduced when they exist on the respective protein sequence. Moreover is important to notice that imidazole used for IMAC protein elution absorbs at 280 nm influencing the protein concentration measurements.

Table 3.7. Summary of the protein properties studied in this work. For E7 CR3 and CBP-ID4 domains, also the values of protein including tags are reported in parenthesis, since they can help during the purification steps. The column on the right includes a fast calculation of the molar protein concentration (μM) considering a solution containing 1.0 $mg^{-1}ml^{-1}$ of pure protein (after removal of tags when applicable).

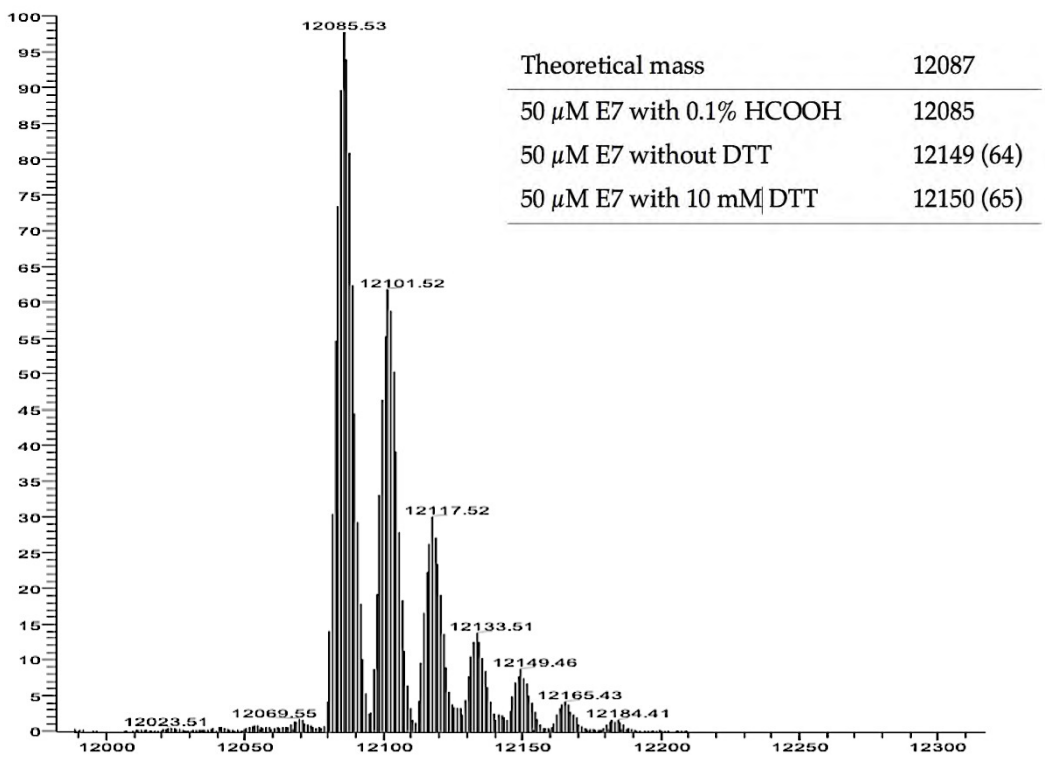
Protein	Nº of residues	MW (Da)	pI	ε ($M^{-1}cm^{-1}$)	μM (1 mg/ml)
HPV16 E7	**106**	**12087**	**4.7**	**5960**	**83**
HPV16 CR3	**58** (98)	**6401** (11044)	**6.0** (7.1)	**1490** (8940)	**156**
HAdV 2/5 13S E1A	**299**	**33176**	**4.9**	**7450**	**30**
HAdV 2 12S E1A	**243**	**26500**	**4.4**	**1490**	**38**
CBP-ID4	**211** (242)	**22539** (26051)	**12.1** (11.4)	**6990** (8480)	**46**

MASS SPECTROMETRY

The potential of mass spectrometry (MS) to analyze proteins is due to the advances gained through the development of soft ionization techniques such as electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI) that can transform biomolecules into ions. ESI can efficiently be interfaced with separation techniques expanding the range of applications in life sciences (Di Marco and Bombi, 2006). MALDI, has the advantage of producing singly charged ions of peptides and proteins, minimizing spectral complexity. Regardless of the ionization source, the sensitivity of a mass spectrometer is related to the mass analyzer where ion separation occurs. Both quadrupole and time of flight (ToF) mass analyzers are commonly used and they can be configured together as QToF tandem mass (known as MS/MS) spectrometric instruments (El-Aneed *et al*., 2009). Recently non-denaturing ESI-MS has been used to study IDPs (Testa *et al*., 2013).

Unlabeled proteins were used for mass spectrometry as well as for other biophysical techniques like SAXS, CD and DLS. A 50 μM HPV16 E7 stock sample was prepared in 10mM Hepes pH 7.5, 150mM KCl, 10μM $ZnCl_2$, 10mM DTT was dialyzed against degassed 10mM ammonium acetate pH 6.6

buffer, using sealed containers. Mass spectra were recorded in positive reflectron mode of a MALDI-TOF mass spectrometer (Bruker Daltonics Ultraflex TOF/TOF). The protein mass data files were analyzed afterwards using the PERL script algorithm. For ESI-MS an E7 aliquot from the 50μM stock solution in 10mM ammonium acetate pH 6.6 buffer was used. The addition of 0.1% formic acid was needed to enhance protonation and increase sensitivity of ESI experiment. ESI-MS spectra were recorded for 5 min in the m/z region from 500 to 30000 Da. For example, the E7 apoprotein theoretical mass is 12087 Da, which in turn containing 1 Zn atom is 12152. Mass spectrometry measured by ESI analysis showed 12149 Da representing the presence of 1 Zn atom. After treatment with 0.1% formic acid for 2.5 hours ESI mass was 12085 Da representing the apoprotein with released zinc (summarized on Figure 3.3).



Theoretical mass	12087
50 μM E7 with 0.1% HCOOH	12085
50 μM E7 without DTT	12149 (64)
50 μM E7 with 10 mM DTT	12150 (65)

Figure 3.3. The HPV 16 E7 protein ESI mass spectrum, after the treatment with 0.1% formic acid for 2.5h. The highest intensity peak with a molecular weight of 12085 Da represents the E7 apoprotein. The other species with less intensity represents the different free cysteine residues of E7 being oxidized by 1 molecule of oxygen forming Cys-SOH. Also the E7 mass values before treatment with formic acid are described. The values in parenthesis are difference compared to the measured E7 apoprotein. The standard atomic weight of zinc is 65.38 ± 0.002 Da. Values in Dalton (Da).

The HAdV 2/5 13S E1A was also prepared using unlabeled nitrogen and carbon sources. The final E1A protein stock was 10μM prepared in 10mM Hepes buffer at pH 7.5 with 150mM KCl, 10mM DTT and 10μM $ZnCl_2$ and was dialyzed against degassed 10mM ammonium acetate pH 6.6 buffer, using sealed containers. MALDI and ESI experiments were recorded always in the presence of DTT. The theoretical molecular weight of E1A apoprotein is 33176 Da while containing 1 atom of zinc is 33241 Da.

CIRCULAR DICHROISM

Circular Dichroism (CD) is based on the differential absorption of left- and right-handed circular polarized light. CD is a fast technique, which allows for the assessment of the secondary structural properties of a protein or protein regions. IDPs present particular CD characteristic different from those of folded proteins and also different from random coil polypeptides, presenting specific conformational preferences, thus revealing partially populated secondary content. These dynamic secondary structure elements can be stabilized or perturbed by temperature; by different chemical agents like solvent, pH, ionic strength, and reducing agents; by post-translational modifications, like phosphorylation and by the presence of metal ligands. Thus the secondary structural properties of IDPs can be studied by CD measurements changing the chemical conditions to analyze the nature of the intrinsically disordered protein samples.

Proteins present CD bands in the far-UV or amide region (175–250 nm) giving information about the secondary structure content, mostly based on the asymmetric conformation attained by the main polypeptide backbone. CD bands in the far-UV region are characteristic for different types of secondary structure. Alpha helix structure displays the most invariable band pattern: a characteristic spectrum with a positive band at 190 nm and two negative bands at 208 and 222 nm. Beta sheet elements, however, are more variable with a positive band at around 198 nm and a single negative band ranging from 214 to 218 nm, depending on the type of structure. The random coil conformation is characterized by a negative band below 200 nm (Figure 3.4)

(Kelly *et al.*, 2005).

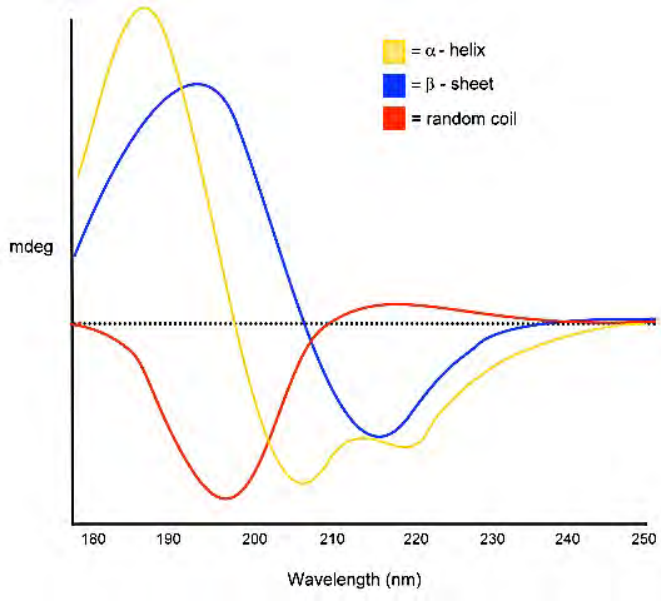


Figure 3.4. CD spectra examples of different pure secondary structures.

The CD measurements were performed in a Jasco J-715 spectropolarimeter using a 0.1 cm path length cell with cooling jacket connected to a water thermostatic device at 25 ºC. Spectra were recorded from 190-250 nm at a scan speed of 20 nm/min combined from an average of 3 scans. The final spectra were baseline-corrected by subtracting the corresponding buffer obtained under identical conditions. Results were shown as the mean residue ellipticity [θ] at a given wavelength. The data were fitted using the secondary structure estimation program K2D3 (Louis-Jeune *et al.*, 2012) [http://www.ogic.ca/projects/k2d3/]. As an example the HPV 16 E7 protein CD were recorded to validate the expected secondary structure prediction of the model. Concentrations from 15μM to 30μM were used and a typical spectrum is reported in Figure 3.5. Its analysis accomplished with the use of the K2D3 webserver confirmed that E7 contains about 50% α-helix and about 10% β-sheet were the remaining 40% is random coil. This data are consistent with bioinformatics predictions indicating that the N-terminus is largely disordered while the CR3 is ordered.

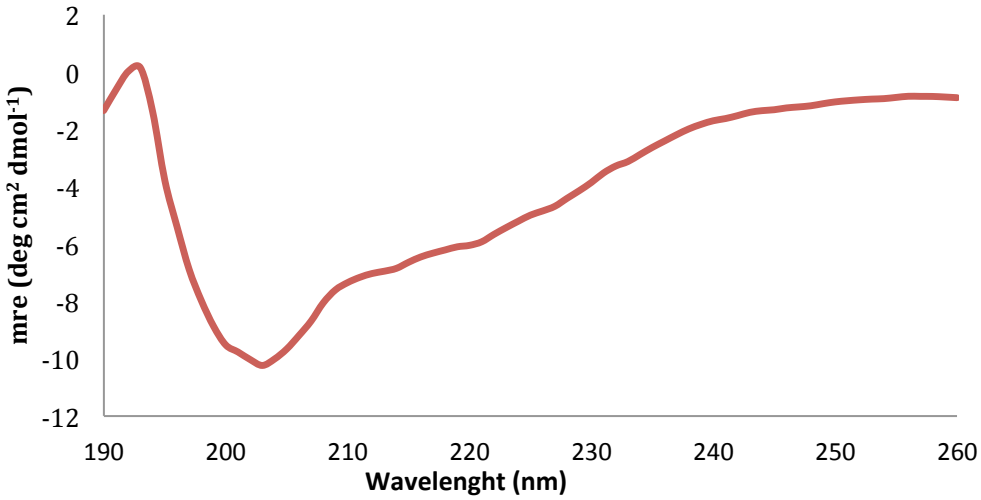


Figure 3.5. CD spectrum of 15μM HPV 16 E7 protein CD spectrum in 5mM KPi buffer at pH 7.5 measured at 25 ºC. The obtained CD spectrum of HPV 16 E7 protein is the typical profile of a protein both containing ordered and disordered regions, was it has been already reported in the literature (Garcia-Alai *et al*., 2007).

DYNAMIC LIGHT SCATTERING

Molecular parameters such as size, molar mass, and intermolecular interactions, which are important to identify and characterize IDPs, can be obtained from light scattering measurements.

All techniques based on scattering like dynamic light scattering (DLS), small angle X-ray scattering (SAXS) and small angle neutron scattering (SANS), require very homogeneous samples. The presence of even small fraction of the aggregated material is known to dramatically affect the scattering profile, making interpretation of data difficult.

DLS analyzes the temporal fluctuations of the light scattering intensity caused by hydrodynamic motions in solution. Thus, DLS is an appropriate technique to determine the hydrodynamic radius (R_H) of a protein also known as the Stokes radius. Measuring the R_H it reflects the apparent size adopted by the solvated tumbling molecule and thus is possible to monitor expansion or compaction of protein molecules. This is especially important for IDPs, which can be recognized and characterized by comparing the measured R_H radii with those calculated for particular reference states, such as the compactly folded or the fully unfolded states.

In the present thesis DLS was performed using a SEC column attached to DLS instrument to avoid the presence of aggregates and to separate multispecies in solutions like monomer and multimers. Samples in a concentration range of 0.1-1.0 mM were injected on a Superdex75 HR-10/30 size exclusion column on a AKTA FPLC system (Amersham Pharmacia Biosciences) connected with a multiangle light scattering (DAWN-EOS, Wyatt Technologies, Santa Barbara, CA) coupled with a quasielastic light scattering detectors. Data analysis and MW were automatically calculated using Wyatt's Astra software. For example, the HPV 16 E7 protein DLS measurements were performed using the same buffer used for SAXS and NMR experiments, 10mM Hepes pH 7.5, 150mM KCl, 10μM $ZnCl_2$ in presence of 10mM DTT. The E7 sample was centrifuged for 5min at 13000 rpm at 4ºC and 100 μl from the supernatant with a concentration of 150μM was injected a Superdex75 HR-10/30 size-exclusion column at a flow-rate of 0.6 ml/min and analyzed as described above. Using these conditions, the DLS as well the refractive index measurements versus elution volume are represented on Figure 3.6.

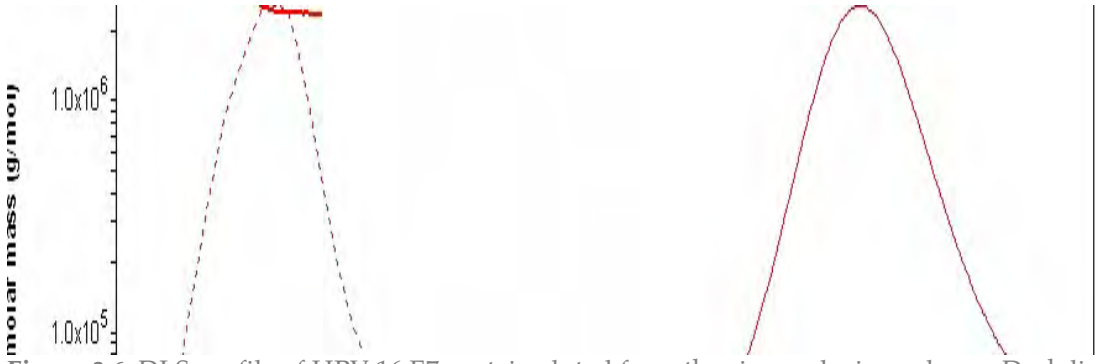


Figure 3.6. DLS profile of HPV 16 E7 protein eluted from the size-exclusion column. Dash line is the scattering light and line is the refractive index. The first peak showed an R_H of 250 Å and the second peak showed an R_H of 15 Å. The chromatogram also shows the molecular mass estimation, but it was not considered in this work because in DLS this calculation is based in the partial volume considering a spherical protein, which is not the case.

The R_H measured was 15 Å meaning an elongated extent of the 106 residues of the analyzed E7 protein in solution. Is important to notice that DLS molecular weight estimation is calculated on the basis of the empirical equation $M_r = 4/3N_A(R_H\ f/f_0)^3/V_k$, where N_A is Avogrado's number, f/f_0 is the ratio of frictional coefficients (set to 1.2 for spherical proteins), and V_k is the partial volume set to 0.73 for a spherical protein.

SMALL ANGLE X-RAY SCATTERING

It is becoming evident that IDPs are not fully disordered, but they have all sort of transient, short, and long-range structural organization that are function-related. The major techniques toward describing the resultant structural ensemble made use of structural calculations supported by NMR and small-angle X-ray scattering (SAXS) data. SAXS, mostly contributes information on the hydrodynamic behavior and topology of the polypeptide chain. In a SAXS experiment, samples containing soluble protein are exposed to an X-ray beam. The different scattered beam intensities are further recorded by a detector as function of the scattering angles of the soluble protein (fig. 3.6), giving rise to an isotropic scattering intensity (I). The solvent scattering is subtracted and the background corrected intensity is presented as a radially averaged one-dimensional curve I(s).

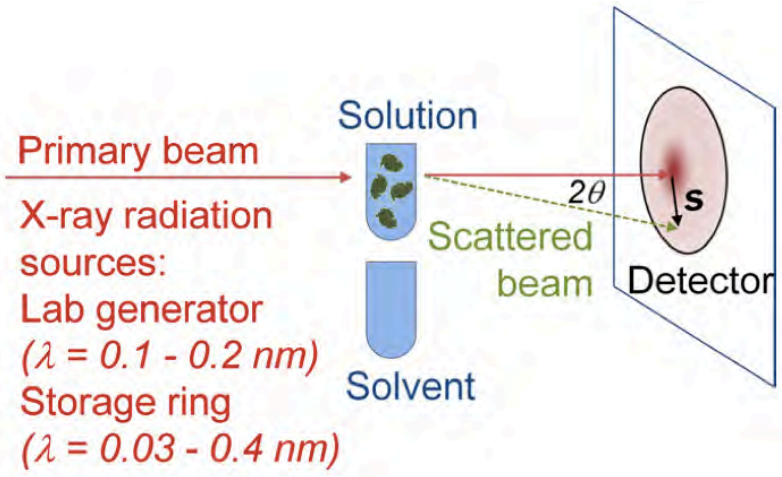


Figure 3.6. Schematic representation of a solution SAXS experiment, adapted from (Petoukhov and Svergun, 2013).

For monodisperse solutions of non-interacting identical and randomly oriented proteins, the SAXS curve is proportional to the scattering of a single particle averaged over all orientations. Thus, the scattering profile carries information about the major geometrical parameters of the particle. Particularly, molecular mass (M) of the solute and its radius of gyration (R_g) are derived from the slope of the Guinier plot. Moreover, the values of the hydrated particle volume (V) and its specific surface (S) can be obtained using the so-called Porod invariant (Petoukhov and Svergun, 2013).

SAXS has been actively used to characterize flexible protein, using advanced approaches that take the insights of conformational flexibility of IDPs (Bernadó and Svergun, 2012a; Bernadó *et al*., 2007a). A set of different conditions to be studied should be designed in advance in order to estimate the number of condition/samples to measure. Conditions such as buffer composition, ionic strength, pH, temperature among others, can be used to analyze the conformational properties of the IDP under investigation. The data analysis of the SAXS measurements can be performed using the available software package like ATSAS, which is in use at DESY (Konarev *et al*., 2006; Petoukhov *et al*., 2012). The following is the workflow implemented at the beam line workstation for the analysis of the data. The standard procedure starts with the subtraction of the buffer for each measurement. Then the SAXS curve is loaded on the program PRIMUS to generate a Kratky plot where the folding state of the protein can be assessed (Figure 3.7) and thus qualitatively identify the IDP conformational disorder (Mertens and Svergun, 2010).

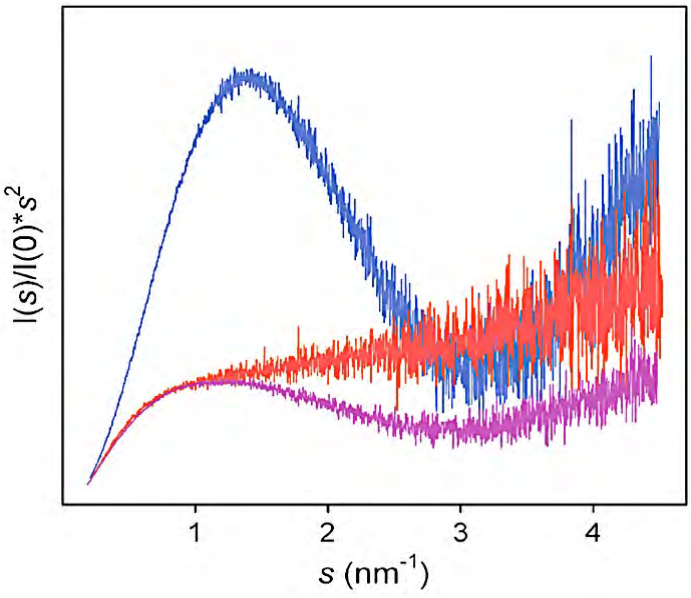


Figure 3.7. Kratky plot representation of typical SAXS curves measured for different protein folding domains. (blue) globular domain; (red) partly disordered domain; (purple) fully disordered domain. Figure adapted from (Bernadó and Svergun, 2012a).

Then the Guinier tool is used to calculate the *Rg* and I(0) by a simple linear fit in logarithmic scale, derived from the very small angles initial part of the scattering profile. Rg of a protein is defined as the root mean square distance

of all the protein atoms from the center of gravity of the molecule. The *I(0)* provides the estimate of the molecular mass (MM) of the protein and hence suggests its oligomeric state. Then the program GNOM is used to calculate the distance distribution function, *P(r)* defined as the distribution of distances between volume elements inside the protein particle. Using the P(r) function it is possible to define the maximum particle distance, *Dmax* (Svergun, 1992).

The SAXS analysis of flexible systems takes in consideration an infinite amount of conformations that are in fast equilibrium exchange. The tool Ensemble Optimization Method (EOM) has been developed as a strategy for the structural characterization of IDPs using SAXS (Bernadó *et al*., 2007a) and is also included in the ATSAS package. The use of EOM approach consists in three consecutive steps: 1) generation of a pool of conformations; 2) selection of a sub-ensemble of conformations describing the SAXS data; 3) quantitative description of the structural properties of the selected ensemble, as discussed in details in (Bernadó and Svergun, 2012b).

The sample preparation for SAXS experiments should follow standard guidelines (Jacques *et al*., 2012) and requires highly pure, monodisperse proteins in a concentration range from 1mg -10mg/ml to fulfill the condition of a "dilute" solution. The concentrations must be determined with accuracy as is required to appropriately normalize the scattering data and thus to estimate the effective molecular mass of the solute. If the sample is aggregated, the scattering data will be difficult or even impossible to interpret. Depending on the SAXS station used, a typical sample volume of 50μl is required for each single measurement. Each SAXS experiments of different condition such as buffer, ionic strength, pH, temperature etc., requires 1-2 mg of purified sample in order to prepare at least 3 different concentrations. The concentration range can be prepared by dilution of a concentrated stock, if is known the protein is not affected by aggregation. In case the protein tends to aggregate it is better to prepare a stock of diluted protein sample and perform the concentrated samples, right before the SAXS experiments. Is important to highlight that for each experiment at a given condition, the scattering of the buffer is also measured by SAXS for further subtraction, thus the background can be corrected and the intensity is

presented as a radially averaged one-dimensional curve I(s). The buffer composition must precisely match the composition of the sample. A small mismatch in the chemical composition of the solvent between the buffer and the sample may lead to difficulties during background subtraction. Thus, the best approach is to use the dialysis buffer where the protein was prepared.

In the present work we used SAXS for the characterization of HPV 16 E7. This protein was prepared in 10mM Hepes pH 7.5, 150mM KCl, 10mM DTT, 10μM $ZnCl_2$ as a diluted stock solution of 50μM (0.6 mg ml^{-1}). The samples where concentrated right before the SAXS measurements to prepare a concentration range from 1.3 - 2.7 mg ml^{-1} (108 - 223 μM). The results are reported in chapter 4.2.

Nuclear Magnetic Resonance Spectroscopy

Due to the high flexibility and heterogeneous nature of IDPs/IDRs, NMR is the only spectroscopic technique that can provide atomic resolution structural and dynamic information on these macromolecules. Together with SAXS, NMR has been the key tool to provide experimental information on the importance of disorder and flexibility for protein function. The atomic resolution experimental results are going to contribute to the understanding not only of the individual characteristics of IDPs, but also of their behavior when interacting either with other protein partners or with possible drugs to accomplish their function. The highly flexible nature of IDPs induces extensive conformational averaging, reducing the nuclear chemical shift dispersion (Bertini *et al*., 2012). Taking flexibility to its extreme, chemical shifts progressively collapse to those of random coil polypeptides, causing extensive resonance overlap. The intrinsic chemical shift dispersion increases from protons to heteronuclei (^{13}C, ^{15}N). Therefore exclusively heteronuclear NMR experiments based on ^{13}C direct detection has crucial relevance for IDPs studies (Bermel *et al*., 2006a; 2006b; Braun *et al*., 1994; Zhang *et al*., 1997) of course in combination with ^{1}H detectable experiments as when studing complex systems, all the information available is welcome. In addition the determination of ^{15}N relaxation rates provides accurate information on the

motional properties of the backbone for each amino acid, as well it provides a general estimation of the expected transverse relaxation rates, which of course have a large impact on the overall sensitivity of multidimensional NMR experiments.

Sequence specific assignment

The sequence specific assignment of the protein backbone resonances is the initial step of NMR experiments to allow getting further structural and dynamic information on IDPs. The most suitable nuclear spins to be exploited for sequence specific assignment are backbone ones, which are more influenced by the contributions of neighboring amino acids (Schwarzinger *et al*., 2001), and carbonyl carbon nuclei and nitrogen nuclei linked in the peptide bond (Braun *et al*., 1994; Dyson and Wright, 2001; Zhang *et al*., 1997). Thus, C´-N correlations display the most favorable chemical shift dispersion allowing the detection of proline residues, often very abundant in IDPs. The chemical shifts can be calculated for any protein by simply taking the primary sequence and using random coil values typical of each amino acid, properly corrected using the contributions from neighboring amino acids. The suite of NMR experiments generally used for protein sequence-specific assignment includes experiments that provide additional correlations involving backbone C´ and N atoms. These experiments exploit the small $^{3}J_{C'C'}$ (Grzesiek and Bax, 1993; Liu *et al*., 2000) as well as multiple magnetization transfer steps mediated by heteronuclear scalar couplings ($^{1}J_{C'N}$, $^{1}J_{C_{\alpha}N}$, $^{2}J_{C_{\alpha}N}$) (Bermel *et al*., 2006b; 2009) to detect correlations with the previous and following carbonyls and nitrogens (Figure 3.8). Another approach to increase the resolution is to move to experiments with higher dimensionality than 3D. Including additional indirect dimensions is now feasible due to improvements in instrumental sensitivity and techniques to reduce the number of sampled points, being very appropriate to increase the resolution study of IDPs (Bermel *et al*., 2012).

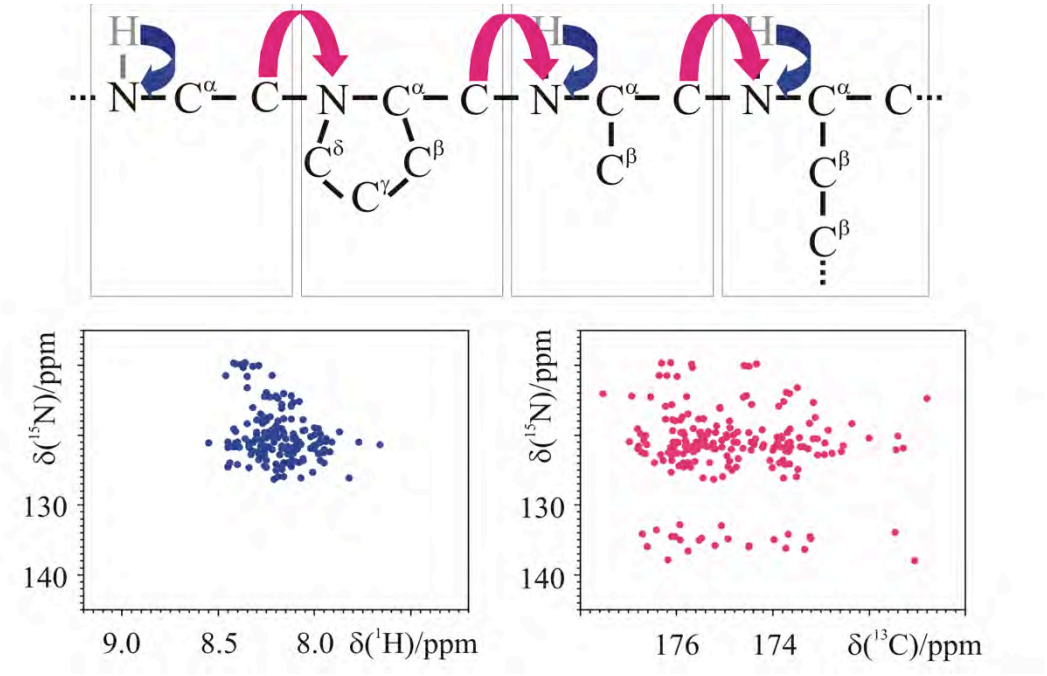


Figure 3.8. Diagram representing an example of sequence specific assignment based on HSQC and CON experiments. The dots in the graphs represent the expected correlations involving backbone nitrogens with the directly bound amide proton (N^{H}_{i}-H^{N}_{i}) or carbonyl carbon (N^{H}_{i}-C'_{i-1}), calculated using random coil chemical shifts appropriately corrected for the contributions from the primary sequence, taking human securin as an example. For a fair comparison, the same spectral widths (in units of Hz) are shown in the two graphs (2 ppm for ^{1}H correspond to 8 ppm for ^{13}C). The HN HSQC and CON experiments acquired on intrinsically disordered human securin are reported in the original publication. Adapted from (Bertini *et al.*, 2012).

Information from NMR chemical shifts

The chemical shifts can be analyzed to understand the secondary structural propensity of the protein. If the sequence-specific assignment of the protein has been achieved it will be possible to attribute the secondary structural propensity to specific regions of the polypeptide chain of an IDP. Because IDPs are characterized by a continuous spectrum of possible disordered states, deviations from random coil chemical shifts can be used to estimate the secondary structural propensities. Compact conformations typical of α-helices and the elongated ones typical of β-strands do cause chemical shift variations in opposite directions for the various nuclei (Spera and Bax, 1991). Therefore, when chemical shifts are identified for several consecutive residues, they can provide information on the secondary structural propensities of different regions of the polypeptide (Marsh *et al.*, 2006).

Information from ^{15}N relaxation rates

^{15}N relaxation rates (^{15}N R_1, ^{15}N R_2, ^{1}H-^{15}N NOEs) provide very useful information to characterize the different local mobility of different parts of the polypeptide chain (Kay *et al.*, 1989; Peng and Wagner, 1994). The heteronuclear NOEs in particular, are the most sensitive reporter of fast local motions on IDPs. Depending on the local effective correlation time, span a large range from -4 to +1 and for this reason provide very effective information on the regions of the protein characterized by different mobility. While a lot of information is directly available from chemical shifts and ^{15}N relaxation rates, additional observables such as exchange effects with the solvent, scalar couplings, residual dipolar couplings, ^{1}H-^{1}H NOEs and PREs can be determined to describe the structural and dynamic properties of the IDP conformational ensemble.

Performing NMR experiments

The most suitable NMR experiments to evaluate the feasibility of a complete NMR characterization of the protein are the 2D experiments correlating the amide nitrogen with the directly bound amide proton (HSQC) or carbonyl (CON). The low chemical shift dispersion of CON provides the first indication that the protein is not characterized by a stable 3D structure. The heteronuclear correlations involving backbone nitrogen nuclei, both for HN and C'N, can be detected with several different variants of the experiments that may help to increase the sensitivity, increase the resolution, decrease experimental time, increase the number of detected cross peaks and so on. The spectral quality can often also be improved through minor changes in the experimental conditions (temperature, pH, buffer, salt, etc.). The suite of experiments generally used for the sequence-specific assignment of folded proteins can also be applied to IDPs, taking care to optimize the experimental set-up for resolution using high number of acquired data points. The long magnetization transfer pathways and the small scalar couplings, which generally drastically reduce the sensitivity of these experiments when applied to study folded proteins, have less impact when used to IDPs since the high protein flexibility causes an increase in coherences lifetime. Thus,

experiments with long coherence transfer delays as well as with multiple coherence transfer steps can be planned. The list of the most suitable 3D NMR experiments, either based on ^{1}H detection or on ^{13}C detection, as well as the correlations expected in each experiment, is reported in Table 3.8. The necessary number of experiments of course depends on the complexity of the protein. A larger set of 3D spectra reduces the time necessary for the assignment and may help to solve ambiguities (Bertini *et al*., 2012).

Table 3.8. Suite of 2D and 3D experiments either based on ^{1}H or on ^{13}C direct detection to achieve protein sequence-specific assignment (only correlations involving backbone nuclei and C^{β} are indicated). Adapted from (Bertini *et al*., 2012).

Experiment	Correlations observed	J couplings
HN HSQC/HMQC	$N(i)\text{-}H^{N}(i)$	$^{1}J_{NH}$
CBCA(CO)NH	$C^{\beta}(i\text{-}1)\text{-}N(i)\text{-}H^{N}(i)$, $C^{\alpha}(i\text{-}1)\text{-}N(i)\text{-}H^{N}(i)$	$^{1}J_{NH}$, $^{1}J_{NC\alpha}$, $^{2}J_{NC\alpha}$, $^{1}J_{NC'}$, $^{1}J_{C\beta C\alpha}$, $^{1}J_{C'C\alpha}$, $^{1}J_{CH}$
CBCANH	$C^{\beta}(i\text{-}1)\text{-}N(i)\text{-}H^{N}(i)$, $C^{\alpha}(i\text{-}1)\text{-}N(i)\text{-}H^{N}(i)$, $C^{\beta}(i)\text{-}N(i)\text{-}H^{N}(i)$, $C^{\alpha}(i)\text{-}N(i)\text{-}H^{N}(i)$	$^{1}J_{NH}$, $^{1}J_{NC\alpha}$, $^{2}J_{NC\alpha}$, $^{1}J_{C\beta C\alpha}$, $^{1}J_{CH}$
HNCA	$C^{\alpha}(i\text{-}1)\text{-}N(i)\text{-}H^{N}(i)$, $C^{\alpha}(i)\text{-}N(i)\text{-}H^{N}(i)$	$^{1}J_{NH}$, $^{1}J_{NC\alpha}$, $^{2}J_{NC\alpha}$
HNCO	$C'(i\text{-}1)\text{-}N(i)\text{-}H^{N}(i)$	$^{1}J_{NH}$, $^{1}J_{NC'}$
HN(CA)CO	$C'(i\text{-}1)\text{-}N(i)\text{-}H^{N}(i)$, $C'(i)\text{-}N(i)\text{-}H^{N}(i)$	$^{1}J_{NH}$, $^{1}J_{NC\alpha}$, $^{2}J_{NC\alpha}$, $^{1}J_{NC'}{}^{1}J_{C'C\alpha}$,
HN(CO)CA	$C^{\alpha}(i\text{-}1)\text{-}N(i)\text{-}H^{N}(i)$	$^{1}J_{NH}$, $^{1}J_{NC\alpha}$, $^{2}J_{NC\alpha}$, $^{1}J_{NC'}{}^{1}J_{C'C\alpha}$,
HN(CA)NNH	$N(i)\text{-}\ N(i)\text{-}H^{N}(i)$, $N(i+1)\text{-}\ N(i)\text{-}H^{N}(i)$, $N(i\text{-}1)\text{-}\ N(i)\text{-}H^{N}(i)$	$^{1}J_{NH}$, $^{1}J_{NC\alpha}$, $^{2}J_{NC\alpha}$
(H)CACO	$C^{\alpha}(i)\text{-}C'(i)$	$^{1}J_{CH}$, $^{1}J_{C'C\alpha}$
(H)CBCACO	$C^{\beta}(i)\text{-}C'(i)$, $C^{\alpha}(i)\text{-}C'(i)$	$^{1}J_{CH}$, $^{1}J_{C\beta C\alpha}$, $^{1}J_{C'C\alpha}$
CON	$C'(i\text{-}1)\text{-}N(i)$	$^{1}J_{NC'}$
(H)CBCACON	$C^{\beta}(i\text{-}1)\text{-}C'(i\text{-}1)\text{-}N(i)$, $C^{\alpha}(i\text{-}1)\text{-}C'(i\text{-}1)\text{-}N(i)$	$^{1}J_{CH}$, $^{1}J_{C\beta C\alpha}$, $^{1}J_{C'C\alpha}$, $^{1}J_{NC'}$
(H)CBCANCO	$C^{\beta}(i)\text{-}C'(i\text{-}1)\text{-}N(i)$, $C^{\alpha}(i)\text{-}C'(i\text{-}1)\text{-}N(i)$, $C^{\beta}(i)\text{-}C'(i)\text{-}N(i+1)$, $C^{\alpha}(i)\text{-}C'(i)\text{-}N(i+1)$	$^{1}J_{CH}$, $^{1}J_{C\beta C\alpha}$, $^{1}J_{C'C\alpha}$, $^{1}J_{NC'}$, $^{1}J_{NC\alpha}$, $^{2}J_{NC\alpha}$
COCON	$C'(i)\text{-}C'(i)\text{-}N(i+1)$, $C'(i\text{-}1)\text{-}C'(i)\text{-}N(i+1)$, $C'(i+1)\text{-}C'(i)\text{-}N(i+1)$	$^{1}J_{NC'}$, $^{3}J_{C'C'}$
(H)NCANCO	$N(i)\text{-}N(i)\text{-}C'(i\text{-}1)$, $N(i+1)\text{-}N(i)\text{-}C'(i\text{-}1)$, $N(i\text{-}1)\text{-}N(i)\text{-}C'(i\text{-}1)$	$^{1}J_{CH}$, $^{1}J_{NH}$, $^{1}J_{C'C\alpha}$, $^{1}J_{NC'}$, $^{1}J_{NC\alpha}$, $^{2}J_{NC\alpha}$

An example of these NMR methodologies is shown on figures 3.9 and 3.10, where both E1A 13S and 12S can be compared. The E1A 13S has been assigned and a scientific manuscript is being submitted. The E1A 12S data is under analysis and is going to be compared with the E1A 13S data.

Figure 3.9. ^{1}H-^{15}N HSQC and ^{13}C-^{15}N CON-IPAP spectra recorded for HAdV 2/5 E1A13S. Experiments were recorded at 16.4 T and 278 K for a 0.2 mM E1A13S sample in 10 mM HEPES buffer at pH 7.5, 50 mM KCl, 10 µM $ZnCl_2$, and 10 mM DTT.

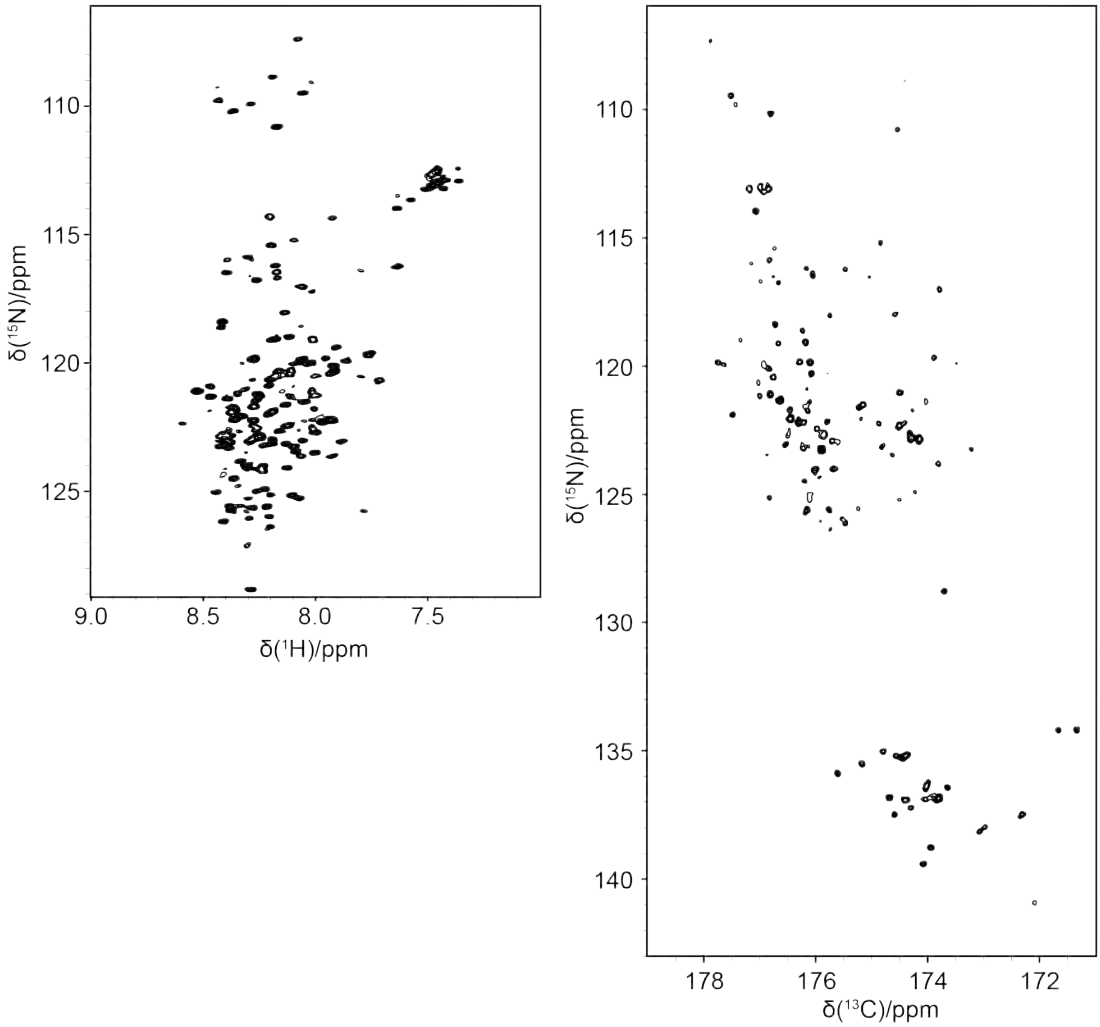
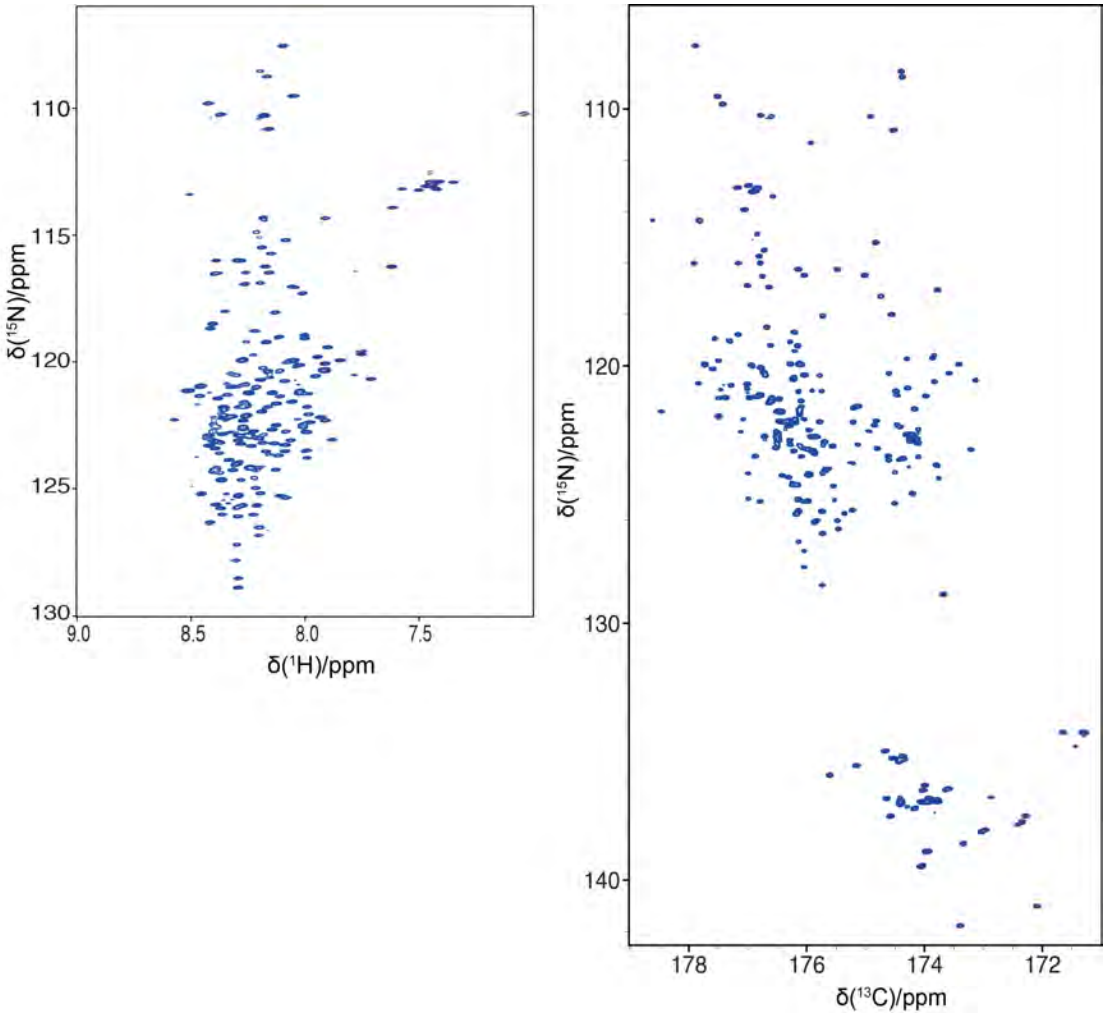


Figure 3.10. ^{1}H-^{15}N HSQC and ^{13}C-^{15}N CON-IPAP spectra recorded for HAdV 2 E1A12S. Experiments were recorded at 16.4 T and 278 K for a 0.8 mM E1A12S sample in 10 mM HEPES buffer at pH 7.5, 50 mM KCl, 10 µM ZnCl2, and 10 mM DTT.



4. Results

NOTE: The author of the present work performed the molecular biology, biochemical and biophysical analysis and characterization of all studied protein constructs presented. He carried out the protein characterization using light scattering, circular dichroism, small-angle X-ray scattering and simple NMR experiments. During the PhD course he was seconded for a period at VIB, Brussels, under the supervision of Prof. Peter Tompa, to learn different techniques of protein expression and purification of IDPs. He also spent a period at DESY, Hamburg, under supervision of Prof. Dmitri Svergun, to learn how to perform SAXS experiments and be able to analyze the SAXS data. In all the projects he participated to data analysis and manuscripts' writing.

4.1.

The Heterogeneous Structural Behavior of E7 from HPV16 Revealed by NMR Spectroscopy

Eduardo O. Calçada[a], Isabella C. Felli[a, b], Tomáš Hošek[a] and Roberta Pierattelli[a, b]

[a] Magnetic Resonance Center (CERM), University of Florence Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy

[a] Department of Chemistry "Ugo Schiff", University of Florence Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

CHEMBIOCHEM
FULL PAPERS

ChemPubSoc Europe

DOI: 10.1002/cbic.201300172

The Heterogeneous Structural Behavior of E7 from HPV16 Revealed by NMR Spectroscopy

Eduardo O. Calçada,[a] Isabella C. Felli,*[a, b] Tomáš Hošek,[a] and Roberta Pierattelli*[a, b]

This paper is dedicated to the memory of our Maestro, Prof. Ivano Bertini

The E7 protein from human papillomavirus (HPV) plays a key role in oncogenesis; for this reason, it is a target of great biomedical interest. To date, no high resolution information is available for the full protein. We present here the NMR characterization of the entire E7 from HPV16, one of the most oncogenic variants of the virus. The protein is very heterogeneous in terms of structural and dynamic properties with a highly flexible N-terminal module and a more structured C terminus. This opens possibilities for studies of molecular-level interactions and post-translational modifications of the protein to unravel functional details that might be linked to its highly oncogenic potential.

Introduction

The importance of local flexibility in determining the function of proteins was recognized long ago and has been widely scrutinized.[1-5] If the extent of local flexibility is taken to extremes, it leads to completely random coil behavior of polypeptide chains, indicated as intrinsic disorder,[6-8] through a wide variety of intermediate cases, both in terms of extent of mobility and in terms of protein stretches involved. Intrinsically disordered proteins (IDPs) possess complementary properties to well-folded ones. Many examples of IDPs have appeared in the literature, showing how their structural plasticity, adaptability, and intrinsic flexibility can actually be key features that enable them to interact with a variety of different partners, adapt to different conditions, and act as hub proteins.[9,10] These peculiar properties provide functional advantages to IDPs, enabling them to play key roles in many regulatory processes.[11] Their function/malfunction has also been related to several diseases.[12]

Thanks to the largely extended and exposed backbones, IDPs often display short amino acid segments (short linear motives, SLiMs) that might interact with different partners.[13,14] This strategy of exploiting proteins' SLiMs for driving interactions, seems to be extensively used by viruses to infect healthy cells. In virtue of their small genomes, able to code only a limited number of proteins, they need economic ways to interfere with the host.[13] This is the case of human papilloma virus (HPV).[15]

HPV is a family of more than 100 different viruses.[16] They can be clinically classified as "low risk" (LR), causing benign warts, and as "high risk" (HR), which are causative agents of cervical cancer, the second most common cancer in women worldwide.[17] The entire proteome of HPV includes six non-structural proteins (E1, E2, E4, E5, E6, and E7) and two structural proteins (L1 and L2). The oncogenic potential of HR HPV relies on the expression of E6 and E7, which are consistently present in HPV-associated carcinomas and are necessary for induction and maintenance of the transformed phenotype.[18-20]

E7 is a small protein comprising about 100 amino acids. Three conserved regions (CR), namely CR1 and CR2 in the N-terminal half and CR3 in the C-terminal half,[21-23] have been identified (Figure 1). The first two (CR1 and CR2), which have been found to interact with a high variety of target proteins,[19] show a high propensity to be disordered.[24] The CR3 region, which contains two CXXC motifs separated by 29 to 30 amino acids, has been shown to bind zinc and is expected to be more structured.[24]

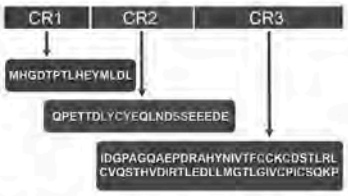


Figure 1. Schematic representation of HPV16 E7 oncoprotein amino acid sequence, and division into three conserved regions.

[a] E. O. Calçada,[+] Prof. I. C. Felli, T. Hošek,[+] Prof. R. Pierattelli
Magnetic Resonance Center (CERM), University of Florence
Via Luigi Sacconi 6, 50019 Sesto Fiorentino (Italy)
E-mail: felli@cerm.unifi.it
pierattelli@cerm.unifi.it

[b] Prof. I. C. Felli, Prof. R. Pierattelli
Department of Chemistry "Ugo Schiff", University of Florence
Via della Lastruccia 3, 50019 Sesto Fiorentino (Italy)

[+] These authors contributed equally to this work.

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/cbic.201300172.

and for the HPV45 variant through nuclear magnetic resonance (NMR) spectroscopy.[26] However, the whole protein failed to give crystals and/or was not characterized through NMR. Atomic resolution information for HPV16 E7 is still lacking,[26] despite the interest in accessing high resolution information for the protein responsible for the high oncogenic potential of this variant and its interactions.

Among the many documented interactions of E7 with host proteins,[15] the one with the retinoblastoma tumor suppressor (pRB) plays a key role in the oncogenic potential of E7 and has been extensively characterized by many techniques.[27] The structure of a short polypeptide containing the LXCXE motif localized in the CR2 region[28] and bound to pRB is available, providing precious hints toward understanding how this short fragment binds to pRB.[29] However, this polypeptide has much lower affinity for pRB with respect to the whole protein,[28] indicating that other parts of the protein also contribute to binding.[30] E7 binds to several other partners, increasing their activity to enhance pRB degradation. The cullin 2 ubiquitin ligase complex is bound by the CR1 of the E7 protein.[31] The CR3 also interacts with the subunit 4 (S4) ATPase of the S26 proteasome and enhances its APTase activity, providing another pathway for pRB degradation.[32] E7 also interacts through its N terminus with other pRB-related tumor-suppressor proteins, p107 and p130.[33,34] pRB degradation is just one example of the complex array of interactions established by viruses like HPV to hijack cell regulation and ensure survival.

To contribute to the understanding of the sophisticated mechanisms behind HPV-related diseases, it is of primary importance to access high resolution structural and dynamic information of the small, though complex, oncogenic proteins of HPV. In this work, we present the characterization by NMR of the full construct of E7 from HPV16. Despite a small molecular mass, its heterogeneous nature limited the possibility of accomplishing the complete sequence-specific assignment but provided clear hints regarding its structural and dynamic features.

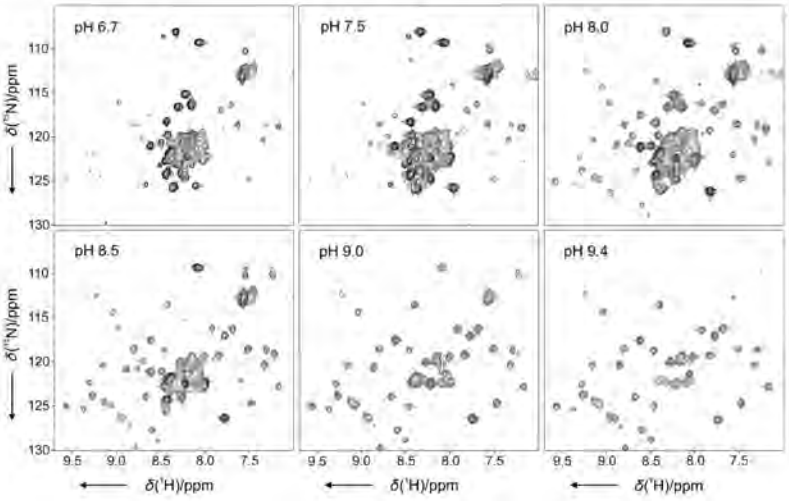


Figure 2. Portions of the ^{1}H–^{15}N SOFAST-HMQC spectra recorded for HPV16 E7 at pH 6.7, 7.0, 7.5, 8.0, 8.5, and 9.4. Spectra have been processed to maximize signal-to-noise at the expense of resolution. The higher signal-to-noise ratio for the set of strong signals clustered in the central region was obtained at pH 7.5. The second set of signals, characterized by a larger chemical shift dispersion, is most intense at pH 8.5. Spectra were recorded at 18.8 T and 298 K for a 0.3 mM sample in 10 mM HEPES buffer, 150 mM KCl, 10 μM $ZnCl_2$, and 10 mM DTT.

Results

The heterogeneous structural and dynamic properties of HPV16 E7 are evident from the 2D ^{1}H–^{15}N correlation NMR spectra reported in Figure 2. The figure shows two distinct sets of signals with very different properties in terms of chemical shift dispersion, signal-to-noise ratio (S/N), and line width. In the first set, clustered in a very narrow spectral region (7.9–8.6 ppm in the ^{1}H dimension), signals are characterized by relatively sharp lines and high S/N and tend to disappear with increasing pH. These are all distinctive properties of signals belonging to nuclei of a highly flexible, disordered, and solvent-exposed module. Elevated exchange rates of amide protons with solvent, typical of unfolded polypeptides, justify the disappearance of the signal with increasing pH, broadening lines beyond detection as observed in the present case.

A second set of much weaker and broader signals, characterized by a larger chemical shift dispersion (7.0–9.6 ppm in the ^{1}H dimension), becomes detectable with increasing pH (Figure 2, pH 8.0–9.4). This second set of signals becomes dominant with increasing pH, whereas the first set decreases in intensity and eventually becomes undetectable (Figure 2, pH > 9.0). The large chemical shift dispersion of this second set of signals is a clear indication that they belong to residues of a more structured part of the protein, also consistent with amide protons in a more solvent-protected conformation, which remain detectable up to pH 9.8. However, line widths of the signals are much larger with respect to what is expected for a well-structured, well-behaved, globular protein of small size. Therefore, we ascribe this second set of signals to a more structured part of the protein which does experience extensive line broadening. The extent of disorder of different parts of the protein from different variants of the virus, as predicted through bioinformatic tools, has been linked to their oncogenic potential.[24]

Sequence-specific assignment of the first set of signals is shown in Figure 3. This was obtained through a combination of ^{1}H- and ^{13}C-detected multidimensional NMR experiments, which enabled us to unambiguously assign 95% of the resi-

CHEMBIOCHEM
FULL PAPERS
www.chembiochem.org

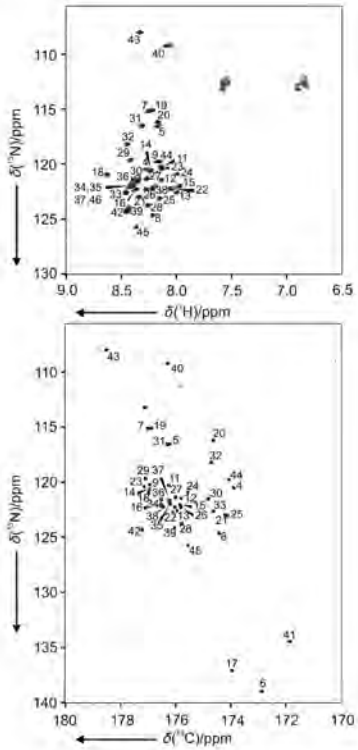


Figure 3. ^{1}H–^{15}N SOFAST-HMQC and ^{13}C–^{15}N CON-IPAP spectra recorded for HPV16 E7. The resulting assignment is reported in the figure, indicating the residue number (relative to the backbone nitrogen). The complete chemical shifts table is reported in Table S1. Experiments were recorded at 16.4 T and 298 K for a 0.76 mM E7 sample in 10 mM HEPES buffer at pH 7.5, 150 mM KCl, 10 μM $ZnCl_2$, and 10 mM DTT.

dues in the N terminus of the protein (E7N, residues 1–46) comprising CR1 and CR2. Chemical shifts of H^N, N, C′, C^α, and C^β of the assigned residues are reported in Table S1 of the Supporting Information and were deposited in the BMRB. The assignment, originally achieved at pH 7.5, was repeated at pH 8.5 to evaluate possible chemical shift changes of the already assigned crosspeaks and to try to detect a sufficient number of correlations involving the second set of crosspeaks to enable their sequence-specific assignment. However, whereas on one hand it was possible to easily transfer the assignments of the crosspeaks from pH 7.5 to pH 8.5, the sensitivity of the triple resonance experiments was not sufficient to perform the sequence-specific assignment of the second set of crosspeaks identified in 2D ^{1}H–^{15}N maps, even though several variants were tested, including those that were relaxation optimized.

The second set of signals could also, in principle, be due to a second conformation of E7N. To exclude this possibility and confirm that the second set of observed signals derives from the final part of the protein comprising CR3, a shorter E7 construct (residues 45–98, E7C) was expressed and purified. The recorded 2D ^{1}H–^{15}N correlation map of E7C (Figure S1) confirmed that the second set of signals observed in analogous spectra recorded on the full construct was due to the final module of the protein (45–98, E7C). This comparison also shows that general spectral characteristics of E7C alone and of E7C in the full construct are very similar.

Chemical shifts are the first reporters of the structural and dynamic properties of a protein that can be analyzed after sequence-specific assignment. Indeed, analysis of experimental heteronuclear chemical shifts (N, C′, C^α, and C^β) and comparison with the so-called random-coil chemical shift values can be used to investigate local secondary structural propensities and ascertain to what extent a certain local conformation is populated. The secondary structure propensity (SSP) score,[35] which combines the different chemical shifts into a single residue-specific parameter, is shown in Figure 4.

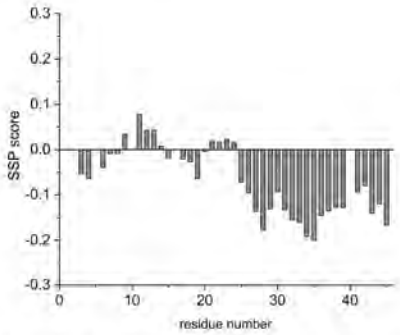


Figure 4. Secondary structure propensity (SSP) score,[35] obtained by combining N^H, C′, C^α, and C^β chemical shifts for the assigned resonances of HPV16 E7.

To have a more general picture of the structural and dynamic properties of the protein through NMR, additional observables were determined. The most relevant were the ^{15}N R_1, ^{15}N R_2, and ^{1}H–^{15}N NOEs, which can be used to identify parts of the polypeptide chain characterized by different motional properties. The magnitude of the heteronuclear ^{1}H–^{15}N NOE effects is one of the most informative measures, as it depends on the local effective correlation time and is very sensitive to fast local motions. Transverse relaxation rates are also sensitive to exchange processes. The combined analysis of these data can be used for characterization of the protein. These data, measured at pH 7.5, are reported in Figure 5 both for the assigned E7N regions (left panels) as well as for the unassigned E7C regions (right panels). The latter are ordered according to increasing ^{15}N R_2 values to facilitate comparison.

All the NMR data determined for the E7N region indicate that it is highly flexible and highly disordered, as is also immediately evident from the SSP values as well as from negative ^{1}H–^{15}N heteronuclear NOE. Detailed inspection of these data also indicates short fragments of amino acids characterized by a mild secondary structural propensity, starting at residue 26 onwards (Figure 4). From a general point of view, the deter-

CHEMBIOCHEM
FULL PAPERS
www.chembiochem.org

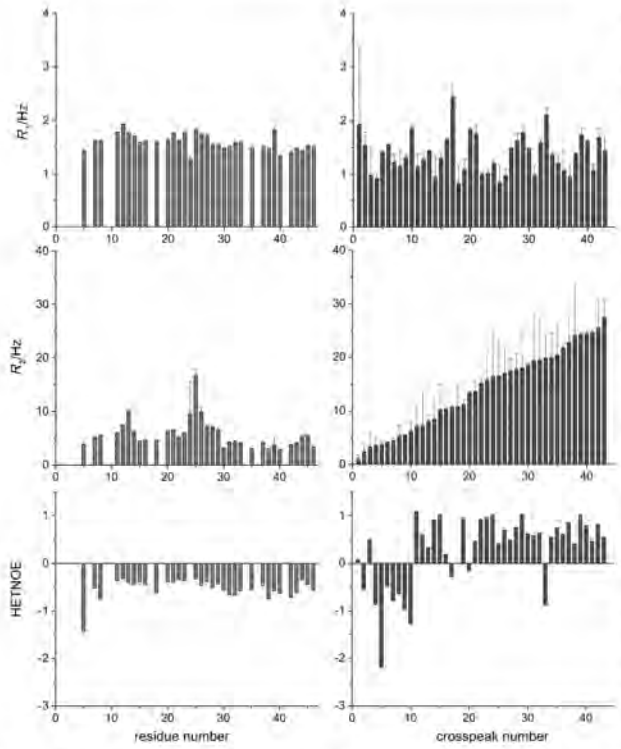


Figure 5. Relaxation data for backbone amide ^{15}N nuclei of HPV16 E7 as a function of the residue number for the N-terminal region (E7N) and of the peak number for the C-terminal region (E7C), with the latter ordered according to increasing R_2 values. From top to bottom: ^{15}N longitudinal relaxation rates (R_1), ^{15}N transverse relaxation rates (R_2), and ^{1}H–^{15}N NOE values (HETNOE). Experiments were recorded at 11.8 T and 298 K for a 0.25 mM sample in HEPES buffer at pH 7.5, 150 mM KCl, 10 μM $ZnCl_2$, and 10 mM DTT.

deriving from a "loose", molten globule-like structure, or from the tendency of the protein to aggregate and form high molecular mass species in solution in equilibrium with monomeric or dimeric forms. All structural evidence available on the constructs of the C terminus from different viral types shows that the CR3 region tends to form dimers,[25,26] and biophysical data support the tendency of the CR3 to form oligomers.[37,38] The latter tendency might be also responsible for the disappearance of the CR3 peaks at low pH.[37,38] In order to gain some insights into possible origins of the observed broadening involving the C terminus of the protein, concentration-dependent measurements and temperature-dependent measurements were collected.

Concentration dependence was monitored through a series of 2D ^{1}H–^{15}N correlation maps at pH 7.5 (Figure 6). This figure reports the ratios of intensities of selected crosspeaks, accurately chosen to be representative of the E7N as well as of the more structured E7C regions of the protein. From the average intensity ratio, a function of the protein concentration for the two regions of the protein (right panel), it is clearly evident that the intensity of signals in E7N increases linearly with the concentration, whereas for the other region, this ratio increases much more slowly with the concentration in agreement with the hypothesis of the formation of high molecular mass species that involve the E7C, leaving the E7N highly flexible and thus still detectable by NMR.

The temperature dependence of the selected crosspeaks shows that the signals deriving from the E7C irreversibly disappear when the temperature is increased above 35 °C, in agreement with the formation of high-order aggregates that are not detectable by NMR spectroscopy. Indeed, upon increase in temperature and loss of the E7C signals, no precipitate is observed at either

mined data agree with the overall picture of a protein characterized by a very flexible N terminus and a more structured C terminus which, however, is characterized by transverse relaxation rates that are much larger than expected for a protein of about 50 amino acids. Calculation of ^{15}N relaxation rates through the HYDRONMR program,[36] by using the monomeric (PDB ID: 2EWL) and dimeric (PDB ID: 2F8B) CR3 NMR structures of the HPV45 E7[26] variant as input (see the Supporting Information), indicates the presence of dimeric forms of CR3. Furthermore, some residues in CR3 display even higher R_2 value than expected for a dimeric form. This could be due to conformational exchange processes

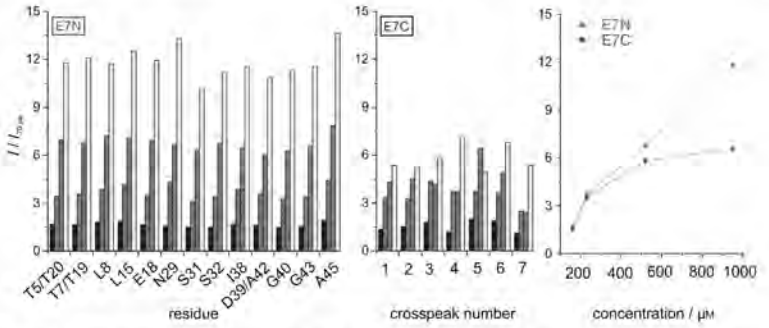


Figure 6. Diagram of the intensity of selected ^{1}H–^{15}N SOFAST-HMQC signals for HPV16 E7 recorded at various concentrations of protein: 160, 230, 520, and 950 μM. Data are scaled to the intensity of the signals from a 70 μM sample. The two sets of signals were selected to be representative of the E7N (left) and of the E7C (right) regions of the protein. In the right panel, the average intensity ratio as a function of the protein concentration for the two parts of the protein are compared. The experiments were recorded at 18.8 T and 298 K for samples in 10 mM HEPES buffer at pH 7.5, 50 mM KCl, 10 μM $ZnCl_2$, and 10 mM DTT.

CHEMBIOCHEM
FULL PAPERS
www.chembiochem.org

pH 7.5 or 8.5, likely due to the fact that E7N at this pH is highly charged and responsible for the high solubility of this protein. The set of signals from E7N is still detectable at temperatures above 35 °C, even at high temperature as observed for other disordered proteins.[39]

Many different experimental conditions were tested through 2D ^{1}H–^{15}N NMR spectra in order to evaluate possible changes in the observed crosspeaks upon changes in ionic strength, buffer, and type and concentration of reducing agent, as well as to improve the S/N and overall relaxation properties of E7C signals. However, the overall behavior of the protein was similar and in line with the presence of a highly flexible N-terminal region, indicating the stability of this protein element under a variety of different conditions. On the other hand, the signals deriving from the C-terminal region of the protein were detectable under specific conditions and were very sensitive to increasing temperature but also resistant to denaturation (not shown).

Discussion

The NMR data obtained for HPV16 E7 are in line with a protein characterized by very heterogeneous structural and dynamic behavior, with a highly flexible N-terminal module (E7N) and a more structured C-terminal region (E7C). The N terminus comprises the so-called CR1 and CR2, as identified by comparison of the primary sequences of E7 from different types of viruses. This part of the protein does contain a few short linear motifs, which are involved in interactions with protein partners. These include the LXCXE motif (22–26), identified as the key motif for interaction with pRB,[28,29] and the SS motif (31–32), the known site of phosphorylation by casein kinase II.[40] Atomic resolution information for this part of the protein is not available for any virus variant; thus, the current data can provide detailed information useful for characterizing the structural and dynamic properties of this part of the polypeptide sequence that, until now, has been broadly defined as disordered. Inspection of NMR data confirms that the N-terminal region is characterized by high flexibility, as evident from the negative values of ^{1}H–^{15}N NOEs and chemical shift values near those of random-coil. The ^{1}H–^{15}N NOEs do approach zero, progressing with the amino acid sequence, and eventually become positive as they approach the C-terminal region of the protein, indicating a less flexible behavior.

The part of the polypeptide sequence from residues 26–46 results from NMR data characterized by a mild propensity (~10%) to adopt an elongated structured conformation. This is the fragment immediately following the LXCXE motif and containing the SS phosphorylation site. It is interesting to note that the LXCXE motif in a nine-residue peptide mimicking the E7 interacting sequence has been shown to bind pRb in a β-strand-like conformation.[28] The propensity to adopt an elongated conformation, identified here for the residues immediately following this fragment, could be relevant to promote binding and also include the phosphorylation site.

The highly charged E7N, responsible for the high solubility of the protein near neutral pH, could also be responsible for preventing the interaction of the N-terminal with the C-terminal regions of E7, maintaining E7N in a fairly elongated form to promote interaction with partners and to ensure that the kinase motif is available for post-translational modifications. Moreover, keeping the two modules apart might be an important feature to render them "independent" one from the other but still synergistic, in order to provide two independent units to modulate the interaction with other partners but that remain next to each other. The availability of the sequence-specific assignment of E7N opens the path to atomic resolution characterization of its interaction with partners.

On the other hand, as stated above, no assignment could be obtained for the C-terminal part of the protein. However, a large set of crosspeaks from this part of the protein was detected in ^{1}H–^{15}N correlation experiments and can be used to determine overall structural and dynamic properties of this module of the protein. The large chemical shift dispersion of the signals is a clear indicator of a more structured polypeptide which, however, is very dissimilar to a folded protein of similar size, due to its markedly different relaxation properties. An increase in transverse relaxation of this module, monitored both from ^{1}H and ^{15}N line widths as well as by ^{15}N transverse relaxation rates, could arise from either conformational exchange broadening, due to a loosely defined tertiary or quaternary structure which affects nuclear relaxation rates but is not loose enough to affect chemical shifts, or by formation of higher molecular mass species in equilibrium in solution, or by an interaction with the N-terminal flexible module. ^{15}N relaxation rates, in particular for ^{1}H–^{15}N NOEs, indicate that the structure is not fixed but that it is characterized by a significant degree of local motion, in agreement with the presence of large local fluctuations that would indicate an initial molten globule state. The concentration dependence of the different sets of signals also indicates aggregation of the C-terminal region while retaining a very flexible N-terminal region. This would provide an interesting mechanism to increase the effective concentration of the N-terminal region by using the C-terminal region to group different molecules together, leaving the N terminus free to move and thus interact with potential partners. The only hypothesis that does not seem to be relevant is that broadening of the C-terminal region is due to mutual crosstalk between the N terminus and C terminus, as the N-terminal region should also be affected in this case. Instead, as mentioned above, the N-terminal region of the protein under conditions close to physiological conditions, seems to provide the driving force for the high solubility of the protein and does remain structurally disordered under a variety of different experimental conditions.

Detection of the signals of the C-terminal region of the protein, even in the absence of their sequence-specific assignment, might also enable us to detect possible structural changes involving this part of the protein upon interaction.[23,41] Indeed, it has been shown that it is important to evaluate the interactions of E7 with other proteins in its full form, as different components appear to have complementary roles, and shorter constructs display lower affinities.[18,29,42] Further work is in progress through different NMR-based strat-

CHEMBIOCHEM
FULL PAPERS
www.chembiochem.org

egies to extend assignment to the C-terminal region of the protein.

Conclusions

The thorough NMR investigation of E7 from HPV16 presented here reveals very heterogeneous structural and dynamic properties, presenting a variable extent of disorder which could be important for the function of the protein, well beyond the accepted schematic view of CR1 and CR2 being fully disordered and CR3 well-structured. In addition, this NMR study provides an important platform for the study of the post-translational modifications and interactions of E7 from HPV16 with the many partners known to interact with different parts of the protein. In conclusion, E7 from HPV16, constituted of <100 amino acids, is able to efficiently interact with several key protein partners, interfering with crucial cell-regulatory processes. The large structural and dynamic heterogeneity monitored here by NMR spectroscopy might be an important aspect in this respect, and the current study opens the possibility for atomic resolution characterization of these interactions, which could provide important information to understand the features that promote progression of cells to malignancy[20,43] and possibly to design new drugs.[44]

Experimental Section

Construction of bacterial expression vectors: The recombinant vector pET20-E7, containing the E7 gene from HPV16, was generously provided by Scil Proteins GmbH.[45] As previously described,[45] the gene was modified to generate silent mutations in codons for His2, to eliminate the possibility of alternative translation initiation and premature termination of translation, and for Pro6, to increase the protein expression yield (pET20-E7-opt). The optimized plasmid construction was transformed into *E. coli* XL1-Blue competent cells (Stratagene), and the resulting clones were screened by DNA sequence analysis of miniprep-purified plasmid samples.

The DNA encoding HPV16 E7 CR3 (45–98 aa) was amplified by PCR from the pET20-E7 plasmid encoding the full-length protein and inserted into the Gateway pEntr-TEV-d-Topo vector to create an entry clone. The entry clone containing the gene of interest was transferred into N-terminal fusion pDEST17 vector containing a His_6 tag by a Gateway LR Clonase II recombination reaction to create the expression clone pDEST17-E7CR3.

Proteins expression and purification: The recombinant vector pET20-E7-opt was transformed into *E. coli* BL21(DE3)pLysS (Stratagene) for protein expression. One colony was selected to inoculate LB medium (50 mL) containing ampicillin and chloramphenicol grown overnight at 37 °C and 180 rpm. For ^{15}N- and $^{13}C,^{15}N$-labeled protein production, a preinoculum (10 mL) was used to inoculate minimal medium (1 L, 48.5 mM Na_2HPO_4, 22.0 mM K_2HPO_4, 8.5 mM NaCl, 0.2 mM $CaCl_2$, 2.0 mM $MgSO_4$, 10.0 μM $ZnCl_2$, 1 mg L^{-1} each of biotin and thiamin, 7.5 mM $(^{15}NH_4)_2SO_4$, and 11.1 mM glucose/$^{13}C_6$-glucose) at 37 °C with constant agitation at 160 rpm. When the OD_{600} reached 0.6, cells were induced with IPTG (1 mM final concentration) and allowed to grow for an additional 4 h at 30 °C. The cells were harvested by centrifugation at 8000 *g* for 20 min, and the cell pellet was stored at −20 °C.

Frozen cells were thawed and suspended in 40 mL of equilibration buffer A (50 mM NaH_2PO_4, 300 mM NaCl, and 10 mM imidazole, pH 8.0). Cells were disrupted by sonication on ice (at 40% sonication power) with cycles of 3 s with 10 s delay pulses for 10 min.

Lysed cells were centrifuged at 165000 *g* for 30 min at 4 °C, and the supernatant was loaded on a 5 mL Ni^{2+} HiTrap column (Amersham Biosciences), pre-equilibrated with buffer A under anaerobic conditions. The column was washed with 10 column volumes of buffer A containing imidazole (50 mM). The E7 protein was eluted with buffer A (40 mL) containing imidazole (250 mM) and then dialyzed against HEPES (10 mM) and KCl (50 mM, pH 7.5). Samples were concentrated to 3.0 mL in the presence of dithiothreitol (DTT; 10 mM) and were stored at 4 °C for biophysical analyses. Mass spectrometry (MALDI and ESI) was performed to confirm the molecular mass of purified E7 protein. All samples for NMR experiments were prepared in HEPES buffer (10 mM) containing KCl (150 mM), $ZnCl_2$ (10 μM), and DTT (10 mM).

For the CR3 construct, the vector was expressed in BL21 (DE3)-pLysS cells, and expression was conducted as for the full-length construct. After harvesting and lysing the cells, the protein remained in the insoluble phase as inclusion bodies. The pellet was suspended in buffer B (8 M urea, 5 mM imidazole, 5 mM DTT, 10 μM $ZnCl_2$, and 10 mM HEPES, pH 8.5) and loaded on Ni^{2+} HiTrap column pre-equilibrated with same buffer. The column was washed with 10 column volumes of buffer B containing imidazole (10 mM). The E7 CR3 fused protein was eluted with buffer B containing imidazole (50 mM) and concentrated to 500 μL and a concentration of 30 mg mL^{-1} for further refolding. Refolding was performed inside a glove box by drip dilution at one drop every 30 s into 50 mL degassed buffer C (10 mM HEPES, 50 mM KCl, 10 mM DTT, and 10 μM $ZnCl_2$, pH 8.5) to achieve 100-fold dilution and a final protein concentration of 0.3 mg mL^{-1}. Refolded protein was concentrated to 1 mL and buffer-exchanged into TEV buffer (20 mM Tris pH 8.0, 50 mM KCl, 5 mM citrate, 10 μM $ZnCl_2$, and 0.5 mM TCEP). The His tag was cleaved from the recombinant protein by incubation with TEV protease and was removed by reverse Ni^{2+} affinity chromatography in HEPES (10 mM, pH 7.5) containing KCl (50 mM) and $ZnCl_2$ (10 μM). Cleaved, pure E7 CR3 protein was concentrated to 300 μL with a final concentration of 40 μM.

NMR experiments: For determination of optimal sample conditions, different variants of 2D ^1H–^{15}N experiments (SOFAST-MQC,[46] HSQC[47,48]) were carried out on Bruker DRX 500 and Bruker Avance 700, 800, and 900 instruments, all equipped with cryogenically cooled probe heads.

The ^1H-detected NMR experiments for sequence-specific assignment were acquired at 298 K on the 18.8 T Bruker Avance 800, operating at 800.13 MHz for ^1H, equipped with a cryogenically cooled probe. ^{13}C-detected, exclusively heteronuclear, NMR experiments for sequence-specific assignment were carried out at 298 K on a 16.4 T Bruker Avance 700, operating at 176.03 MHz for ^{13}C, equipped with cryogenically cooled probe optimized for ^{13}C sensitivity. A dataset of ^1H-detected (SOFAST-HMQC,[46] HNCO,[49,50] HNCACO,[51] CBCACONH,[52] HNCACB,[53] and HN(CA)NNH[54]) and ^{13}C-detected (protonless CON-IPAP,[55] HCACO-IPAP,[56,57] and HCBCACON-IPAP[57]) NMR experiments was necessary for sequence-specific assignment of H^N, N, C', C^α, and C^β resonances.

To further characterize the structural and dynamic features of HPV16 E7 oncoprotein, the following ^1H-detected NMR experiments were acquired at 298 K: ^{15}N relaxation experiments (R_1, R_2, and ^1H–^{15}N NOE)[5,58] experiments on an 11.7 T Bruker DRX 500

CHEMBIOCHEM
FULL PAPERS
www.chembiochem.org

spectrometer equipped with a cryogenically cooled triple-resonance probe head. Experimental details are reported in Table S2.

Data were processed with TopSpin 2.0 and were analyzed with the programs CARA[59] and Sparky.[60] The secondary structure propensity from chemical shifts was determined by using the SSP program.[35] The ^{15}N relaxation rates (R_1 and R_2) were determined by fitting the cross peak volumes, measured as a function of variable delay, to single-exponential decay. ^{1}H–^{15}N NOE values were obtained as a ratio between peak volumes in spectra recorded with and without ^{1}H saturation. Relaxation rate calculation was performed by using HYDRONMR[36] version 7C.

Acknowledgements

This work was supported in part by the Joint Research Activity and Access to Research Infrastructures (BioNMR, contract 261863) and by the Marie Curie ITN programs (IDPbyNMR, contract 264257) in the EC 7th Framework.

Keywords: heteronuclear NMR · HPV · IDP · viral proteins

[1] G. Wagner, S. Hyberts, J. W. Peng in *NMR of Proteins* (Eds.: G. M. Glore, A. M. Gronenborn), CRC Press, New York, **1993**, pp. 220–257.
[2] G. Wagner, *Curr. Opin. Struct. Biol.* **1993**, *3*, 748–754.
[3] A. Mittermaier, L. E. Kay, *Science* **2006**, *312*, 224–228.
[4] A. Mittermaier, L. E. Kay, *Trends Biochem. Sci.* **2009**, *34*, 601–611.
[5] G. Barbato, M. Ikura, L. E. Kay, R. W. Pastor, A. Bax, *Biochemistry* **1992**, *31*, 5269–5278.
[6] I. C. Felli, R. Pierattelli, P. Tompa in *NMR of Biomolecules: Towards Mechanistic Systems Biology* (Eds.: I. Bertini, K. S. McGreevy, G. Parigi), Wiley-Blackwell, Weinheim, **2012**, pp. 137–152.
[7] P. Tompa, *Trends Biochem. Sci.* **2012**, *37*, 509–516.
[8] B. He, K. Wang, Y. Liu, B. Xue, V. Uversky, A. K. Dunker, *Cell Res.* **2009**, *19*, 929–949.
[9] P. E. Wright, H. J. Dyson, *J. Mol. Biol.* **1999**, *293*, 321–331.
[10] V. N. Uversky, C. J. Oldfield, A. K. Dunker, *J. Mol. Recognit.* **2005**, *18*, 343–384.
[11] M. Fuxreiter, I. Simon, P. Friedrich, P. Tompa, *J. Mol. Biol.* **2004**, *338*, 1015–1026.
[12] V. Uversky, C. J. Oldfield, A. K. Dunker, *Annu. Rev. Biophys.* **2008**, *37*, 215–246.
[13] N. E. Davey, G. Trave, T. J. Gibson, *Trends Biochem. Sci.* **2011**, *36*, 159–169.
[14] P. Tompa, M. Fuxreiter, *Trends Biochem. Sci.* **2008**, *33*, 2–8.
[15] L. B. Chemes, J. Glavina, J. Falvovich, G. de Prat-Gay, I. E. Sánchez, *J. Mol. Biol.* **2012**, *422*, 336–346.
[16] H. U. Bernard, R. D. Burk, K. van Doorslaer, H. zur Hausen, E. M. de Villiers, *Virology* **2010**, *401*, 70–79.
[17] H. zur Hausen, *Nat. Rev. Cancer* **2002**, *2*, 342–350.
[18] M. G. Donà in *Oncogene Proteins: New Research* (Eds.: A. H. Malloy, E. C. Carson), Nova Science, Hauppauge, **2008**, pp. 19–63.
[19] M. E. McLaughlin-Drubin, K. Münger, *Virology* **2009**, *384*, 335–344.
[20] L. Banks, D. Pim, M. Thomas, *Nat. Rev. Cancer* **2012**, *12*, 877–886.
[21] J. R. Gage, C. Meyers, F. O. Wettstein, *J. Virol.* **1990**, *64*, 723–730.
[22] W. C. Phelps, K. Münger, C. L. Yee, J. A. Barnes, P. M. Howley, *J. Virol.* **1992**, *66*, 2418–2427.
[23] B. Todorovic, K. Hung, P. Massimi, N. Avvakumov, F. A. Dick, G. S. Shaw, L. Banks, J. S. Mymryk, *J. Virol.* **2012**, *86*, 13313–13323.
[24] V. N. Uversky, A. Roman, C. J. Oldfield, A. K. Dunker, *J. Proteome Res.* **2006**, *5*, 1829–1842.
[25] X. Liu, A. Clements, K. Zhao, R. Marmorstein, *J. Biol. Chem.* **2006**, *281*, 578–586.
[26] O. Ohlenschläger, T. Seiboth, H. Zengerling, L. Briese, A. Marchanka, R. Ramachandran, M. Baum, M. Korbas, W. Meyer-Klaucke, M. Dürst, M. Görlach, *Oncogene* **2006**, *25*, 5953–5959.
[27] L. B. Chemes, I. E. Sanchez, L. G. Alonso, G. de Prat-Gay in *Flexible Viruses: Structural Disorder in Viral Proteins* (Eds.: V. Uversky, S. Longhi), Wiley, Hoboken, **2012**, pp. 313–346.
[28] N. Dyson, P. M. Howley, K. Münger, E. Harlow, *Science* **1989**, *243*, 934–937.
[29] J. O. Lee, A. A. Russo, N. P. Pavletich, *Nature* **1998**, *391*, 859–865.
[30] L. B. Chemes, I. E. Sanchez, C. Smal, G. de Prat-Gay, *FEBS J.* **2010**, *277*, 973–988.
[31] K. Huh, X. Zhou, H. Hayakawa, J. Y. Cho, T. A. Libermann, J. Jin, J. W. Harper, K. Münger, *J. Virol.* **2007**, *81*, 9737–9747.
[32] E. Berezutskaya, S. Bagchi, *J. Biol. Chem.* **1997**, *272*, 30135–30140.
[33] N. Dyson, P. Guida, K. Münger, E. Harlow, *J. Virol.* **1992**, *66*, 6893–6902.
[34] R. Davies, R. Hicks, T. Crook, J. Morris, K. Vousden, *J. Virol.* **1993**, *67*, 2521–2528.
[35] J. A. Marsh, V. K. Singh, Z. Jia, J. D. Forman-Kay, *Protein Sci.* **2006**, *15*, 2795–2804.
[36] J. G. de La Torre, M. L. Huertas, B. Carrasco, *J. Magn. Reson.* **2000**, *147*, 138–146.
[37] L. G. Alonso, M. M. Garcia-Alai, A. D. Nadra, A. N. Lapeña, F. L. Almeida, P. Gualfetti, G. de Prat-Gay, *Biochemistry* **2002**, *41*, 10510–10518.
[38] L. G. Alonso, M. M. Garcia-Alai, C. Smal, J. M. Centeno, R. Iacono, E. Castaño, P. Gualfetti, G. de Prat-Gay, *Biochemistry* **2004**, *43*, 3310–3317.
[39] P. Tompa in *Structure and Function of Intrinsically Disordered Proteins*, CRC Press, Boca Raton, **2009**.
[40] J. M. Firzlaff, B. Lüscher, R. N. Eisenman, *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 5187–5191.
[41] B. Todorovic, P. Massimi, K. Hung, G. S. Shaw, L. Banks, J. S. Mymryk, *J. Virol.* **2011**, *85*, 10048–10057.
[42] D. R. Patrick, A. Oliff, D. C. Heimbrook, *J. Biol. Chem.* **1994**, *269*, 6842–6850.
[43] C. A. Moody, L. A. Laimins, *Nat. Rev. Cancer* **2010**, *10*, 550–560.
[44] D. Fera, D. C. Schultz, S. Hodawadekar, M. Reichman, P. S. Donover, J. Melvin, S. Troutman, J. L. Kissil, D. M. Huryn, R. Marmorstein, *Chem. Biol.* **2012**, *19*, 518–528.
[45] E. A. Mirecka, R. Rudolph, T. Hey, *Protein Expression Purif.* **2006**, *48*, 281–291.
[46] P. Schanda, E. Kupce, B. Brutscher, *J. Biomol. NMR* **2005**, *33*, 199–211.
[47] L. E. Kay, P. Keifer, T. Saarinen, *J. Am. Chem. Soc.* **1992**, *114*, 10663–10665.
[48] J. Schleucher, M. Schwendinger, M. Sattler, P. Schmidt, O. Schedletzky, S. J. Glaser, O. W. Sørensen, C. Griesinger, *J. Biomol. NMR* **1994**, *4*, 301–306.
[49] S. Grzesiek, A. Bax, *J. Magn. Reson.* **1992**, *96*, 432–440.
[50] J. Schleucher, M. Sattler, C. Griesinger, *Angew. Chem.* **1993**, *105*, 1518–1521; *Angew. Chem. Int. Ed. Engl.* **1993**, *32*, 1489–1491.
[51] R. T. Clubb, G. Wagner, *J. Biomol. NMR* **1992**, *2*, 389–394.
[52] S. Grzesiek, A. Bax, *J. Am. Chem. Soc.* **1992**, *114*, 6291–6293.
[53] D. R. Muhandiram, L. E. Kay, *J. Magn. Reson. Ser. B* **1994**, *103*, 203–216.
[54] R. Weisemann, H. Rüterjans, W. Bermel, *J. Biomol. NMR* **1993**, *3*, 113–120.
[55] W. Bermel, I. Bertini, L. Duma, L. Emsley, I. C. Felli, R. Pierattelli, P. R. Vasos, *Angew. Chem.* **2005**, *117*, 3149–3152; *Angew. Chem. Int. Ed.* **2005**, *44*, 3089–3092.
[56] Z. Serber, C. Richter, D. Moskau, J.-M. Boehlen, T. Gerfin, D. Marek, M. Haeberli, L. Baselgia, F. Laukien, A. S. Stern, J. C. Hoch, V. Dötsch, *J. Am. Chem. Soc.* **2000**, *122*, 3554–3555.
[57] W. Bermel, I. Bertini, V. Csizmok, I. C. Felli, R. Pierattelli, P. Tompa, *J. Magn. Reson.* **2009**, *198*, 275–281.
[58] N. A. Farrow, R. Muhandiram, A. U. Singer, S. M. Pascal, C. M. Kay, G. Gish, S. E. Shoelson, T. Pawson, J. D. Forman-Kay, L. E. Kay, *Biochemistry* **1994**, *33*, 5984–6003.
[59] R. Keller, *The Computer Aided Resonance Assignment Tutorial*, Cantina, Goldau, **2004**, pp. 1–81.
[60] Sparky 3, T. D. Goddard, D. G. Kneller, University of California, San Francisco, **2000**.

Received: March 22, 2013
Published online on August 12, 2013

CHEMBIOCHEM

Supporting Information

The Heterogeneous Structural Behavior of E7 from HPV16 Revealed by NMR Spectroscopy

Eduardo O. Calçada,[a] Isabella C. Felli,*[a, b] Tomáš Hošek,[a] and Roberta Pierattelli*[a, b]

cbic_201300172_sm_miscellaneous_information.pdf

Table 1. Sequence specific assignment of H^N, N, C´, C^α, C^β in HPV 16 E7 at 298 K. Proton resonances were calibrated with respect to the signal of 2,2-dimethylsilapentane-5-sulfonic acid (DSS). Carbon chemical shifts were referred to external 2,2-dimethyl-2-silapentane-5-sulfonic acid (DSS), and to the signal of dioxane at 69.31 ppm as secondary reference. Nitrogen chemical shifts were referenced indirectly to the ^{13}C standard using the conversion factor derived from the ratio of NMR frequencies.[1]

#	Res	HN	N	C´	C^α	C^β
1	MET					
2	HIS					
3	GLY			173.881	45.263	--
4	ASP	8.277	120.44	176.262	54.241	41.051
5	THR	8.159	116.529	172.863	59.819	69.781
6	PRO	--	138.96	177.005	63.336	31.946
7	THR	8.243	115.122	174.411	62.034	69.799
8	LEU	8.206	124.585	176.927	55.419	42.207
9	HIS	8.182	120.201	175.481	56.658	30.77
10	GLU					
11	TYR	8.129	120.234	175.783	58.039	38.417
12	MET	8.128	121.422	175.941	55.522	32.546
13	LEU	7.995	122.589	176.927	55.544	42.362
14	ASP	8.23	120.591	175.831	54.262	41.038
15	LEU	7.972	121.987	177.073	54.837	42.285
16	GLN	8.355	122.313	173.941	53.573	28.73
17	PRO	--	137.091	176.967	63.044	31.954
18	GLU	8.621	120.934	176.871	56.795	30.188
19	THR	8.214	115.068	174.626	61.686	69.784
20	THR	8.159	116.142	174.201	61.977	69.775
21	ASP	8.332	122.949	176.025	54.458	41.046
22	LEU	8.042	122.239	177.147	55.426	42.215
23	TYR	8.128	120.397	175.554	57.969	38.412
24	CYS	7.979	120.913	174.102	58.351	27.603
25	TYR	8.136	123.072	175.791	58.164	38.69
26	GLU	8.269	122.277	175.994	56.601	30.347
27	GLN	8.261	121.354	175.756	55.699	29.312
28	LEU	8.253	123.759	177.055	55.414	42.218
29	ASN	8.41	119.622	174.818	53.065	39.156
30	ASP	8.342	121.476	176.285	54.243	41.328
31	SER	8.3	116.529	174.687	58.347	63.913
32	SER	8.441	118.168	174.622	58.585	63.825
33	GLU	8.449	122.626	176.501	56.582	30.204
34	GLU	8.41	122.095	176.185	56.286	30.459
35	GLU	8.394	121.945	176.489	56.479	30.263
36	ASP	8.355	121.572	176.214	54.245	41.327
37	GLU	8.384	121.632	176.463	56.597	30.201
38	ILE	8.206	122.224	176.022	61.284	38.701
39	ASP	8.428	124.114	176.263	54.245	41.338
40	GLY	8.081	109.191	171.846	44.797	--
41	PRO	--	134.462	177.2	63.053	31.951
42	ALA	8.452	124.265	178.516	52.688	19.049
43	GLY	8.324	107.976	174.064	45.35	--
44	GLN	8.152	119.761	175.552	55.525	29.595
45	ALA	8.355	125.712	177.46	52.355	19.331
46	GLU	8.396	121.677	174.7	56.657	30.199

1

Table 2. NMR experiments and acquisition parameters used for sequence specific assignment of H^N, N, C', C^α, C^β resonances and dynamic characterization of HPV16 E7 at 298 K.

The following parameters were used in all experiments. The ^1H carrier was placed at 4.7 ppm, the ^{15}N carrier at 116.9 ppm. Band selective ^{13}C pulses were given at 173 and 39 ppm to excite or invert C' and $C^{\alpha/\beta}$ spins respectively. The following band-selective pulses were used: 300μs with Q5[2] and time reversed Q5 shapes for C^α and C^{ali} excitation, 220 μs Q3[2] shape for C'/C^{ali} inversion. The experiments for sequence specific assignment were acquired with a Bruker AVANCE 800 spectrometer equipped with an inverse triple-resonance cryogenically cooled probehead (^1H-^{15}N SOFAST-HMQC[3], HNCO,[4;5] HNCACO,[6] HNCACB,[7] CBCACONH,[8] HN(CA)NNH[9]) and with a Bruker AVANCE 700 spectrometer equipped with a triple-resonance cryogenically cooled probehead optimized for direct ^{13}C-detection (CON-IPAP,[10] HCACO-IPAP,[11;12] HCBCACON-IPAP[12]). The other experiments (^{15}N R_1, ^{15}N R_2 and ^1H-^{15}N NOEs) were acquired with a Bruker AVANCE 500 spectrometer equipped with an inverse triple-resonance cryogenically cooled probehead.

Experiments	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]	900* (^{15}N)	1024 (^{13}C)		35	30		32	2.0
HCACO-IPAP[c]	660* (^{13}C)	1024 (^{13}C)		50	32		32	1.4
HCBCACON-IPAP[c]	220* (^{13}C)	64 (^{15}N)	1024 (^{13}C)	60.4	36	50	16	1.2
^1H detected								
^1H-^{15}N SOFAST-HMQC	512 (^{15}N)	1024 (^1H)		22	16		2	0.5
HNCO	128 (^{13}C)	60 (^{15}N)	2048 (^1H)	10	22	13.9	4	1
HNCACO	128 (^{13}C)	60 (^{15}N)	2048 (^1H)	10	22	13.9	16	1
CBCACONH	128 (^{13}C)	60 (^{15}N)	2048 (^1H)	75	22	13.9	8	1
HNCACB	128 (^{13}C)	60 (^{15}N)	2048 (^1H)	75	22	13.9	16	1
HN(CA)NNH	70 (^{15}N)	70 (^{15}N)	2048 (^1H)	25	25	13.9	32	1
Dynamics								
^{15}N T_1[d]	128 (^{15}N)	1024 (^1H)		25	14		32	3.0
^{15}N T_2[d]	128 (^{15}N)	1024 (^1H)		25	14		32	3.0
steady-state heteronuclear ^1H-^{15}N-NOEs[d]	256 (^{15}N)	1024 (^1H)		25	14		168	5.0

[a] number of acquired scans.

[b] Relaxation delay in seconds.

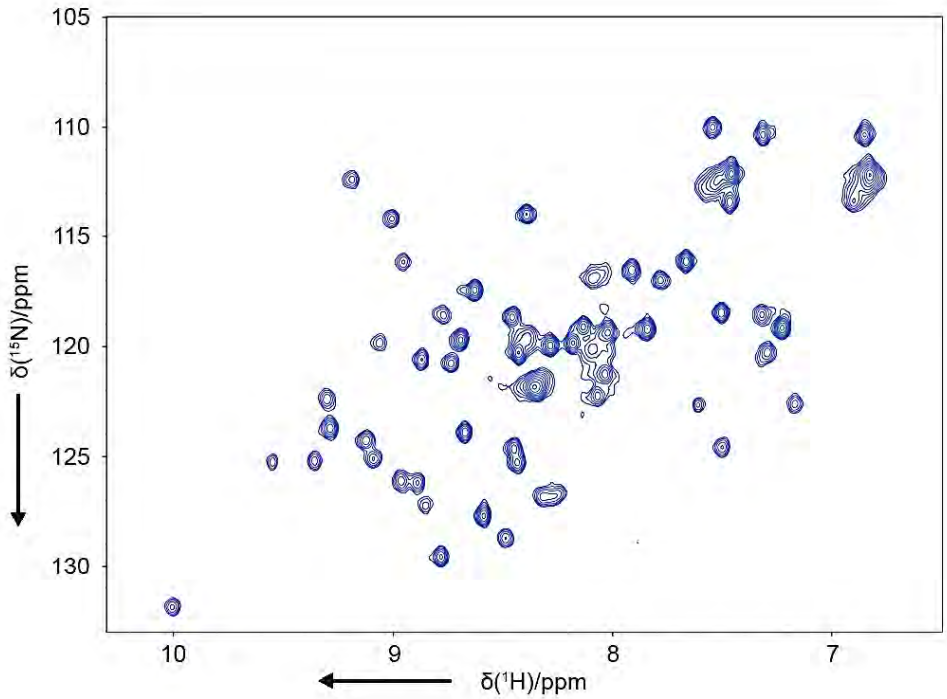
[c] For experiments acquired in the IPAP mode, the dimension in which the two experiments are stored is indicated with an asterisk.

[d] In ^{15}N R_1, ^{15}N R_2 and heteronuclear ^1H-^{15}N-NOEs experiments the water signal was suppressed with 'water flip-back' scheme.[13] For the determination of R_1, 10 experiments were acquired changing the variable delay from 2.5 ms to 1.2 s. For the determination of R_2, 9 experiments were acquired changing the variable delay from 17 to 250 ms.

All 3D and 2D spectra were processed using TopSpin 2.0 (Bruker) and analyzed through the CARA (Computer Aided Resonance Assignment) software developed at ETH-Zürich.

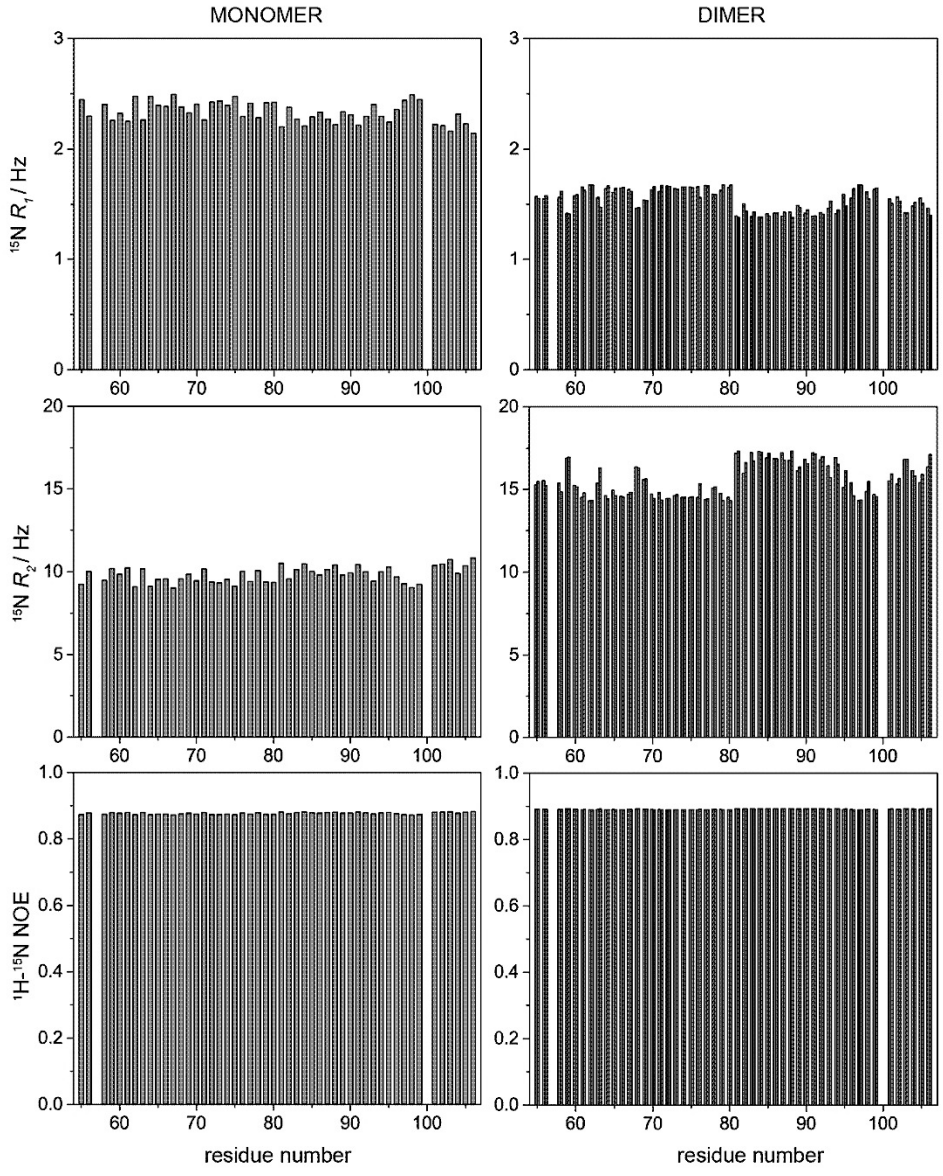
2

Figure S1. The 2D ^{1}H-^{15}N correlation spectrum of the short construct comprising residues 45-98 (E7C) is shown. The spectrum was recorded on a Bruker AVANCE NMR spectrometer operating at 900 MHz and 298 K on a 0.04 mM sample in 10 mM HEPES buffer at pH 7.5, 150 mM KCl, 10 µM $ZnCl_2$ and 10 mM DTT.



3

Figure S2. The ^{15}N relaxation data (^{15}N R_1, ^{15}N R_2 and ^{1}H-^{15}N NOE) calculated with HYDRONMR[14] using as input models the monomeric and dimeric NMR structures of the HPV 45 E7 CR3 (PDB codes 2EWL and 2F8B)[15] are reported below. The data calculated for the monomer are reported in the left panels and those for the dimer in the right ones.



Reference List

[1.] J. L. Markley, A. Bax, Y. Arata, C. W. Hilbers, R. Kaptein, B. D. Sykes, P. E. Wright, K. Wüthrich, *J.Biomol.NMR* **1998,** *12* 1-23.

[2.] L. Emsley, G. Bodenhausen, *Chem.Phys.Lett.* **1990,** *165* 469-476.

[3.] P. Schanda, B. Brutscher, *J.Am.Chem.Soc.* **2005,** *127* 8014-8015.

[4.] S. Grzesiek, A. Bax, *J.Magn.Reson.* **1992,** *96* 432-440.

[5.] J. Schleucher, M. Sattler, C. Griesinger, *Angew.Chem.Int.Ed Engl.* **1993,** *32* 1489-1491.

[6.] R. T. Clubb, G. Wagner, *J.Biomol.NMR* **1992,** *2* 389-394.

[7.] D. R. Muhandiram, L. E. Kay, *J.Magn.Reson.Ser.B* **1994,** *103* 203-216.

[8.] S. Grzesiek, A. Bax, *J.Am.Chem.Soc.* **1992,** *114* 6291-6293.

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

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

[11.] Z. Serber, C. Richter, D. Moskau, J.-M. Boehlen, T. Gerfin, D. Marek, M. Haeberli, L. Baselgia, F. Laukien, A. S. Stern, J. C. Hoch, V. Dötsch, *J.Am.Chem.Soc.* **2000,** *122* 3554-3555.

[12.] W. Bermel, I. Bertini, V. Csizmok, I. C. Felli, R. Pierattelli, P. Tompa, *J.Magn.Reson.* **2009,** *198* 275-281.

[13.] S. Grzesiek, A. Bax, *J.Am.Chem.Soc.* **1993,** *115* 12593-12594.

[14.] J. G. de la Torre, M. L. Huertas, B. Carrasco, *J.Magn.Reson.* **2000,** *147* 138-146.

[15.] O. Ohlenschlager, T. Seiboth, H. Zengerling, L. Briese, A. Marchanka, R. Ramachandran, M. Baum, M. Korbas, W. Meyer-Klaucke, M. Dürst, M. Görlach, *Oncogene* **2006,** *25* 5953-5959.

4.2.

On the super-tertiary structure of HPV 16 E7

[a] Magnetic Resonance Center (CERM), University of Florence Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy

[b] Department of Chemistry "Ugo Schiff", University of Florence Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

[c] EMBL Hamburg, BioSAXS group, Notkestrasse 85, D-22603 Hamburg, Germany

In preparation.

INTRODUCTION

Many intrinsically disordered proteins (IDPs) consist of combinations of globular domains, linear motifs and disordered regions. These IDPs are characterized by a dynamic ensemble of conformations that cannot be easily described by traditional structural biology and have often a heterogeneous structural behavior that render their investigation troublesome.

This is the case of the protein E7 from Human Papilloma Virus (HPV). This protein has been investigated since long time as it is the main transforming product from HPV (Phelps *et al*., 1988) and has been reported to interact with a large number of cellular targets. This feature enables E7 to interfere with multiple cell regulatory pathways, eventually leading to cell transformation and cancer in the case of the most aggressive HPV phenotypes (Moody and Laimins, 2010b).

E7 is constituted by about one hundred amino-acids that can be grouped in three conserved regions (CR), namely CR1 and CR2 in the N-terminal half (E7N) and CR3 in the C-terminal half (E7C) (Phelps *et al*., 1992). The first two conserved regions (CR1 and CR2) show a high propensity to be disordered. The CR3 region, which contains two CXXC motifs separated by 29 or 30 amino acids, has been shown to bind zinc and is more structured (Uversky *et al*., 2006). The structural characterization of the C-terminal part of the protein, which contains the CR3 and the zinc binding motifs, has been achieved for short constructs of the HPV 1 variant through X-ray (Liu *et al*., 2006) and of the HPV 45 variant through nuclear magnetic resonance (NMR) spectroscopy (Ohlenschläger *et al*., 2006). The E7 CR3 of these variants assembles as a dimer. Each subunit of the dimer contains a two-stranded antiparallel sheet formed by residues 44–52 and 58–65 followed by a sharp U-turn leading to 12 amino-acids long helix, a bend and an extended strand, leading to a final short helix. (Liu *et al*., 2006). If the CR3 region is structurally well characterized, the whole protein failed to give crystals and was not characterized through NMR until very recently (Calçada *et al*., 2013).

In the previous work (Calçada *et al*., 2013) we have studied with high-resolution techniques for the first time the full length of E7 from HPV 16, one of the most

relevant variant of the virus, which proved to be responsible for the occurrence of cervix cancer (Hausen, 2002). This protein shares most of the characteristics of E7 isolated from other variants, but resulted to be even more difficult to be investigated due to the high tendency to form macromolecular assemblies (Clements *et al.*, 2000) . We have shown that the combined use of ^{1}H and ^{13}C direct detected NMR experiments on the entire construct of HPV 16 E7 can provide the complete assignment of the first half of the protein, which showed the high mobility and the favorable relaxation properties of an IDP (Calçada *et al.*, 2013) while the second part failed to give useful triple resonance spectra due to large signal broadening and poor sensitivity. The latter effect is likely the result of both the tendency of this protein to form large molecular assemblies and the chemical exchange between monomeric/dimeric/tetrameric forms of the protein (and higher order multimers) (Chinami *et al.*, 1994; Munger *et al.*, 2007). The formation of these multimers is mediated by the CR3 region of HPV E7 but mutations that prevent E7 oligomerization lead to protein instability and disrupt the cell transforming properties of E7 (Fera and Marmorstein, 2012) render the investigation of these mutated forms of the protein far less significant than studies of the wild-type protein.

High-resolution information obtained by NMR spectroscopy can be profitably complemented with data from low-resolution techniques, like small-angle x-ray scattering (SAXS), which are able to provide information of the molecular assembly of a molecule in solution (Bernadó and Svergun, 2012b; Bernadó *et al.*, 2007a). In a SAXS experiment on a dilute macromolecular solution, the measured intensity is the result of the scattering of all the particles in the illuminated sample volume. For monodisperse samples systems containing randomly oriented molecules with identical structures, the isotropic SAXS intensity is proportional to the single particle scattering, averaged over all orientations.

Over the last decade major advances in instrumentation have led to the possibility to investigate an extremely broad range of macromolecules, from small proteins to macromolecular complexes up to several hundred daltons. A concomitant improvement in the computational analysis of the data allow

nowadays to investigate also highly flexible systems such as IDPs. Exploiting the possibility to represent the data by an average of conformers, a pool containing a large number of possible conformations is randomly generated to cover the configurational space and appropriate subsets of models is selected using the "Ensemble Optimization Method (EOM). Even if the interpretation of data obtained for heterogeneous systems, particularly in the presence of polydispersed particles in solution, is challenging this approach, the SAXS data complemented by NMR information are able to provide a description of the supertertiary structure of HPV 16 E7 and of the species present in equilibrium in solution.

MATERIALS AND METHODS

Bioinformatics analysis and modeling

Primary sequences of HPV 16, 1 and 45 were aligned by using the Clustal Omega software tool (Sievers *et al*., 2011). Protein disorder prediction from sequence analysis of these E7 proteins was performed using PONDR-FIT, which detects both the contribution of each residue to the disorder and the disorder content of the protein (Xue *et al*., 2010).

Homology modeling

The E7 sequence from HPV 16 was modeled partially by homology using the NMR solution structure from the Protein Data Bank (PDB) of the C-terminal domain (monomer) of the HPV 45 E7 (PDB ID: 2EWL)(Ohlenschläger *et al*., 2006) which has 65 % similarity with E7 from HPV 16, as a template. The homology modeling was performed using MODELLER 9.12 (Šali and Blundell, 1993).

Protein preparation and purification

The recombinant plasmid pET20-E7 containing the E7 gene from HPV16, was generously provided by Scil Proteins GmbH. As previously described (Ohlenschläger *et al*. 2006) the gene was modified to generate silent mutations in codons for His-2, to eliminate the possibility of alternative translation initiation and premature termination of translation, and Pro-6, to increase the protein expression yield. The optimized and correctly sequenced plasmid construct was transformed into *E. coli* BL21/pLysS competent cells for protein expression. HPV16 E7 were prepared growing cultures at 37ºC, induction at 0.6 optical density (OD) with 1mM IPTG at 30ºC during 4h. Different test expressions were done and best conditions were found obtaining an yield of 36mg/L.

All steps of purification since cell disruption until sample preparation were done in glove-box using degassed buffers containing also 10 mM DTT. Cells were disrupted by sonication and soluble fraction was loaded on a 5ml Ni^{2+} HiTrap column (Amersham biosciences Europe) under anaerobic conditions. The HPV16 E7 protein was washed and eluted with imidazole and finally the buffer was exchanged by using size-exclusion chromatography into 10 mM HEPES, 50 mM KCl, pH 7.5. Samples were concentrated till 3.0 ml in the presence of 10 mM DTT and stored at 4ºC for biophysical analysis.

Dynamic Light Scattering

DLS measurements on E7 protein were done preliminary to SAXS measurements. Protein concentrations were determined with UV and the samples were prepared in 10mM HEPES pH 7.5, 150mM KCl, 10µM $ZnCl_2$, 10mM DTT. Tests without the addition of DTT were also done, establishing the need of the reducing agent in the protein solution.

The E7 sample in a concentration of 150µM was injected on a pre-equilibrated Superdex75 HR-10/30 size-exclusion column at a flow-rate of 0.6 ml/min and at 298 K, controlled by an AKTA FPLC system (Amersham Pharmacia Biosciences). The column was connected to a multiangle light scattering (DAWN-EOS, Wyatt

Technologies, Santa Barbara, CA) coupled with quasielastic light scattering detectors, where the DLS data were recorded. To calculate the R_H and MM the DLS data analysis was performed using Wyatt's Astra software, supplied with the instrument.

Two elution peaks were always observed. The first peak elutes at 6.3 ml, corresponding to an oligomeric form of the protein, with an R_h of 250 Å. The second peak elutes at 9.3 ml with an R_h of 15 Å. Both the analysis of the e refractometer peaks and a native-PAGE showed that the oligomer present in solution was 5 times less concentrated than the dimer, but still sizable. The shapes of the peaks were very large due to the presence of highly flexible regions of E7 protein that enlarge the apparent hydrodynamic radius.

Small-angle X-ray scattering (SAXS)

SAXS data of the HPV 16 E7 protein were collected at the EMBL X33 beamline (DESY, Hamburg) (Roessle *et al*., 2007) using a robotic sample changer (Round *et al*., 2008).

Samples were prepared in 10mM HEPES pH 7.5, 150mM KCl, 10µM $ZnCl_2$, 10 mM DTT. A concentration range from 1.3 - 2.7 mg ml^{-1} (108 - 223 µM) was used. Concentrated samples where prepared right before the measurements using a refrigerated centrifuge with the addition of 10 mM DTT both to keep the protein protected from radiation damage and reduced. Radiation damage during X-ray exposure was monitored by several successive measurements of a single protein solution. No significant changes were observed (data not shown).

SAXS data were recorded on a PILATUS 1M pixel detector (DECTRIS, Baden, Switzerland) at a sample–detector distance of 2.7 m and a wavelength of 1.5 Å. This set up covers a range of momentum transfer of 0.012 b s b 0.6 Å−1 ($s = 4\pi \sin(\theta) / \lambda$, where 2θ is the scattering angle). The measured data were processed by PRIMUS (Konarev *et al*., 2003). The forward scattering *I(0)* and the radius of gyration (R_g) were calculated using the Guinier approximation (Guinier, 1939) and assuming that at very small angles (s b 1.3 / R_g) the intensity is represented

as $I(s) = I(0) * \exp(-(sRg)2 / 3)$. The pair-distance distribution function *p(r)*, from which the maximum particle dimension (D_{max}) is estimated, was computed using GNOM (Svergun, 1992). Finally, the excluded volume of the hydrated particle was computed using the Porod invariant and the molecular mass was estimated as 0.6 times the Porod volume. Low resolution shape analysis of the solute was performed using the *ab-initio* program DAMMIF (Franke and Svergun, 2009). This algorithm represents the macromolecule by an assembly of densely packed beads and employs simulated annealing to build a compact interconnected configuration of beads inside a sphere with a diameter (D_{max}) that fits the experimental data $I_{exp}(s)$ and minimizes the discrepancy $\chi 2$. The protein ensembles based SAXS data where created using the program EOM (Bernadó *et al*., 2007b). All programs used for analysis of the SAXS data belong to the ATSAS package (Petoukhov *et al*., 2012).

RESULTS AND DISCUSSION

Modeling of the structure

The high-risk HPV16 E7 oncoprotein is 38% analogous to the low-risk HPV1a and 45% analogous to the high-risk HPV45. Sequence alignment for all the constructs confined to the CR3 shows that HPV 45 E7 has the highest similarity with the HPV 16 E7, arriving at 65%.

From the comparison among E7 primary sequences by using different structural predictors it results that E7 expected to be highly disordered in the N-terminal of the polypeptide for all the strains here compared, as shown in Figure 1. HPV16 E7 is the one showing the largest tendency of disorder, while HPV1a E7 the least. The CR3 region, instead, is predicted largely structured for all the variants.

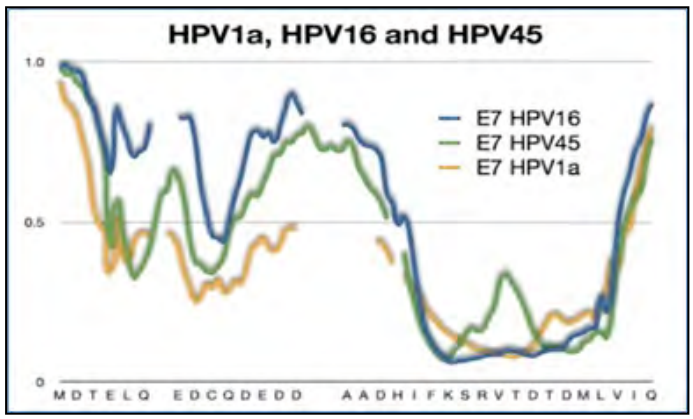


Figure 1. PONDR-FIT prediction tool applied to E7 sequences from HPV1a, HPV16 and HPV45. Values close to 1 indicate high disordered propensity.

The current version of MODELLER 9.12 was used to create a model of the full length HPV 16 E7 protein based on the HPV 45 E7 CR3 monomer (PDB ID: 2EWL) as template. The generated model and the template have a RMSD = 0.496 Å, which is not surprising, seeing the high homology between the primary sequences. In Figure 2 the two structures are superimposed to show the high quality of the fit. The first half of the polypeptide is, of course, randomly placed in the modeled HPV 16 E7 and represents just a screen-shot for the description of its disordered nature.

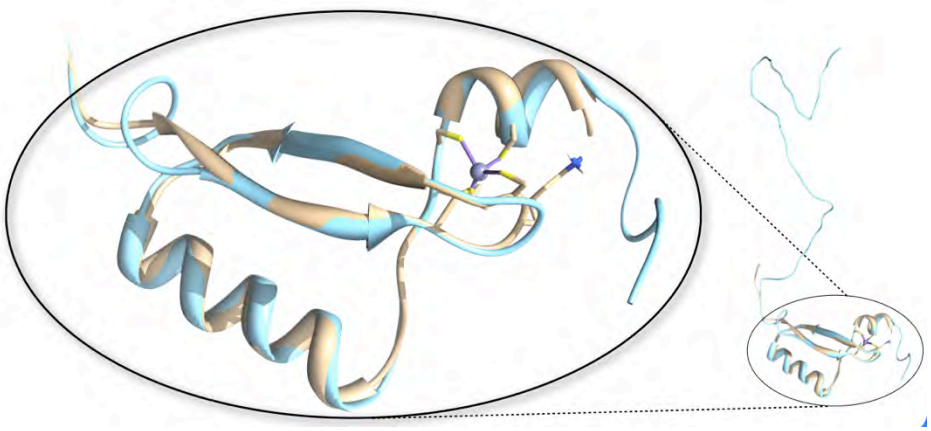


Figure 2. Model of HPV16 E7 (blue) based on 'monomer a' from 2EWL NMR structure of HPV45 E7 protein (brown). Both proteins were created matched and pictures created using the Chimera software. The CR3 region is zoomed to highlight the zinc-binding motif of E7.

SAXS

In Figure 3 the experimental profiles of the SAXS data are reported. The overall parameters derived from the extrapolated curve displayed in Figure 3 are presented in Table 1. The programs DAMMIF and GASBOR were used to compute the *ab-initio* and reconstruction shapes respectively presented in Figure 4.

Based on the Porod volume, the molecular mass of the protein in solution was determined to be about 48 ± 5 kDa, which corresponds well to the expected value for a tetrameric arrangement of the protein, being the monomer about 12 kDa. The *p(r)* function had a shape typical for an elongated particle and revealed a D_{max} = 125 Å. Twelve independent runs of DAMMIF computed without symmetry restrictions were averaged by DAMAVER, providing the molecular mass estimate of about 42 kDa, which is still compatible with a tetramer.

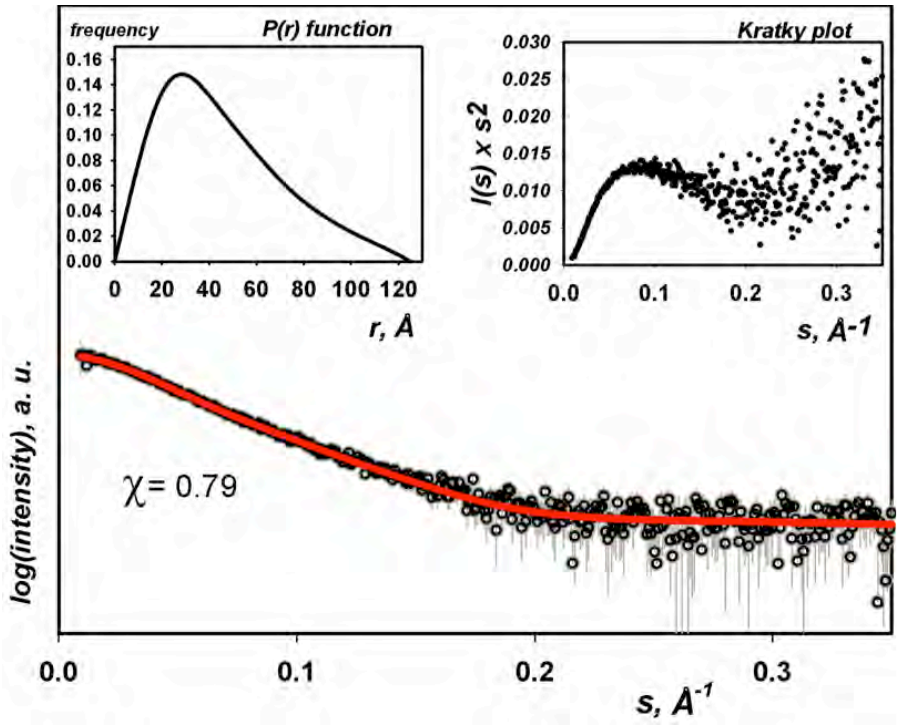


Figure 3. Experimental SAXS data of the E7 oncoprotein and theoretical scattering (red). The logarithm of the scattering intensity is plotted against the momentum transfer, using PRIMUS. The figure also shows the derived pair-distance distribution function *p(r)* and the derived Kractky plot.

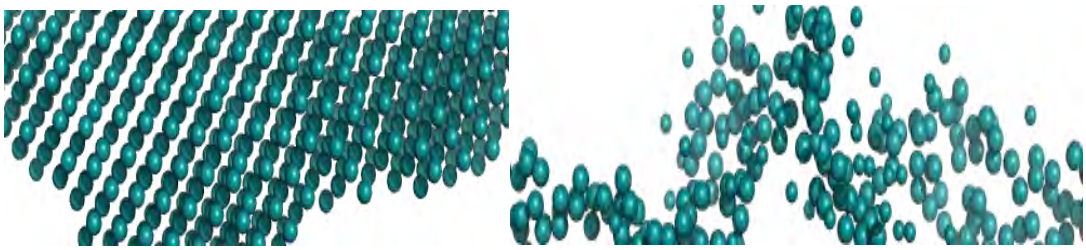


Figure 4. *ab-initio* shape generated with DAMMIF program (left) and reconstruction shape generated with GASBOR (right) data analysis.

Due to the largely unfolded nature of the HPV 16 E7 oncoprotein, the program EOM was used to build a structural model taking into account the flexibility of the system. Since high-resolution information for the structured part of HPV 16 E7 oncoprotein are not available, the model obtained *in-silico* was used as a starting point, taking into consideration the available information on the molecular assembly derived from both the NMR (pdb ID: 2F8B) and the X-ray (pdb ID: 2B9D) structure resolved for the E7 CR3 of HPV 45 and 1, respectively, to define the multimerization interfaces. In this way we assumed the presence of a folded domain and an extended portion of protein, flexible and possibly elongated.

The EOM program generated a pool with 15000 models where multiple runs of the genetic algorithm were performed and the R_g distribution of the selected ensembles compared to the R_g distribution of the entire pool. In this way flexible features of the particle can be speculated. Multiple runs of EOM were performed and the results were averaged to provide quantitative information about the flexibility of the protein in solution (in particular, about the R_g distribution in the selected ensembles). The Figure 5 shows the models with the highest score. These results suggest the presence of tetrameric E7 in solution formed by two dimers, with a P222-like symmetry.

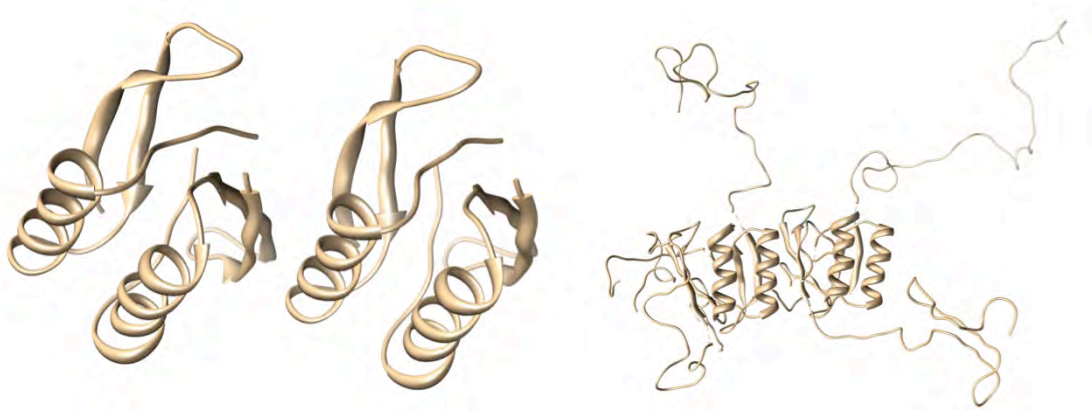


Figure 5. A) Tetrameric E7 generated starting from the modeled monomeric unit assembled taking into account the HPV 45 E7 CR3 and the HPV 1 E7 CR3 structures (pdb ID: 2F8B, 2B9D respectively); B) EOM tetramer model of HPV 16 E7 protein based on the tetrameric unit.

The dimerization involves the first α-helices of each monomer and the formation of an intermolecular two-stranded antiparallel β-sheet between the first β-strand of one monomer and the second β- strand of the other, forming an highly hydrophobic core. The dimers assemble tanks to the interaction established between one α-helices of one dimer and the opposed strand of the other dimer. The disordered regions, as they appear from the EOM models, tend to assume an extended conformation, typical of intrinsically disordered proteins.

However, a careful analysis of the data, very noisy, and of the results of the fitting opens up also the possibility of the presence in solution of a simple dimeric form of the protein. Indeed, a highly extended dimer could result as large as a tetramer according to the R_g. In Figure 6 the possible EOM model obtained assuming only the presence of dimeric E7 cover an ample conformational space.

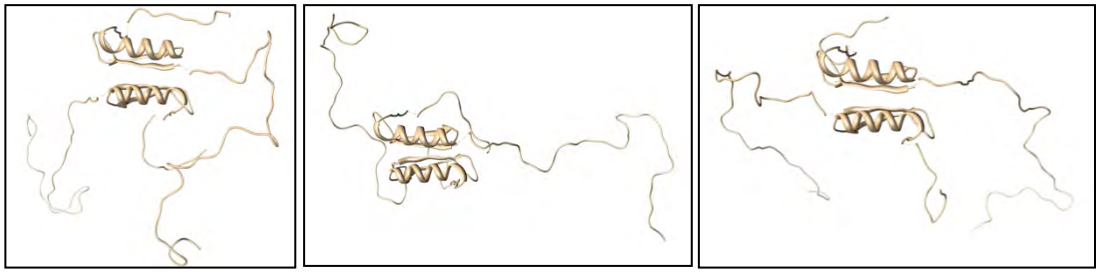


Figure 6. EOM dimeric model of HPV 16 E7 protein.

The presence in solution of different forms of E7 (monomer/dimer/tetramer) has been already proposed based on sedimentation experiments (Clements *et al*., 2000). Also the NMR evidences pointed out the presence of a heterogeneous system, suggesting a high aggregation tendency for the C-terminal part of the protein, still in the presence of an highly mobile and flexible N-terminal part. The presence of polydispersity in solution renders difficult the analysis of the SAXS data. Indeed, the presence of oligomers in solution, even in small amount, produces the scattering of the majority of the X-ray beam and the results are subject of large error. Further studies are in progress to optimize the collection of the SAXS data.

To complete the present study there is the need to implement size-exclusion chromatography directly on the SAXS instrument to be able to separate, at least partially, oligomeric forms of the protein from monomeric, dimeric and tetrameric E7. This part of the work is currently in progress in DESY, Hamburg.

Table 1, SAXS data collection and scattering parameters for HPV 16 E7 protein. Abbreviations: D_{max}: maximal particle dimension; *MM*: molecular mass; R_g: radius of gyration; V_p: Porod volume; V_{ex}: particle excluded volume. The dry volume was calculated using the web tool, http://www.basic.northwestern.edu/biotools/proteincalc.html, (Harpaz *et al.*, 1994).

Instrument	EMBL X33 beam line storage ring DORIS III (DESY, Hamburg)
Beam geometry	2 x 0.6 mm^2
Wavelength (Å)	1.5
q-Range (Å−1)	0.006 - 0.600
Exposure time (min)	2 (8 x 15sec)
Concentration range (mg ml−1)	1.3 - 2.7
Temperature (K)	298
Structural parameters	
$I(0)$ (relative) (from P(r))	11040
R_g (Å) (from P(r))	36 ± 3
$I(0)$ (cm−1) (from Guinier)	11066
R_g (Å) (from Guinier)	36 ± 2
D_{max} (Å)	125 ± 10
Porod volume estimate (Å^3)	69221
Excluded volume estimate (Å^3)	95400
Dry volume calculated from sequence (Å^3)	14734
Molecular mass determination	
Molecular mass *MM* (Da) from Porod volume ($V_p * 0.6$)	41500 ± 5000
Molecular mass *MM* (Da) from excluded volume (V_{ex} / 2)	48000 ± 5000
Calculated monomeric *MM* from sequence	12087
Software employed	
Data processing	PRIMUS
Ab initio analysis	DAMMIF
Validation and averaging	DAMAVER
Computation of ensembles	EOM
3D graphic representations	CHIMERA and MACPYMOL

References

Bernadó, P., and Svergun, D.I. (2012). Analysis of intrinsically disordered proteins by small-angle X-ray scattering. Methods Mol. Biol. *896*, 107–122.

Bernadó, P., Mylonas, E., Petoukhov, M.V., Blackledge, M., and Svergun, D.I. (2007a). Structural characterization of flexible proteins using small-angle X-ray scattering. J. Am. Chem. Soc. *129*, 5656–5664.

Bernadó, P., Mylonas, E., Petoukhov, M.V., Blackledge, M., and Svergun, D.I. (2007b). Structural characterization of flexible proteins using small-angle X-ray scattering. J. Am. Chem. Soc. *129*, 5656–5664.

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

Chinami, M., Sasaki, S., and Hachiya, N. (1994). Functional oligomerization of purified human papillomavirus types 16 and 6b E7 proteins expressed in Escherichia coli.

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

Fera, D., and Marmorstein, R. (2012). Different Regions of the HPV-E7 and Ad-E1A Viral Oncoproteins Bind Competitively but through Distinct Mechanisms to the CH1 Transactivation Domain of p300. Biochemistry *51*, 9524–9534.

Franke, D., and Svergun, D.I. (2009). DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. J. Appl. Crystallogr. *42*, 342–346.

Guinier, A. (1939). La diffraction des rayons X aux tres petits angles: applications a l'etude de phenomenes ultramicroscopiques. Annales De Physique, EDP Sciences.

Harpaz, Y., Gerstein, M., and Chothia, C. (1994). Volume changes on protein folding. Structure *2*, 641–649.

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

Konarev, P.V., Volkov, V.V., Sokolova, A.V., Koch, M.H.J., and Svergun, D.I. (2003). PRIMUS: a Windows PC-based system for small-angle scattering data analysis. J. Appl. Crystallogr. *36*, 1277–1282.

Liu, X., Clements, A., Zhao, K., and Marmorstein, R. (2006). Structure of the Human Papillomavirus E7 Oncoprotein and Its Mechanism for Inactivation of the Retinoblastoma Tumor Suppressor. Journal of Biological Chemistry *281*, 578–586.

Moody, C.A., and Laimins, L.A. (2010). Human papillomavirus oncoproteins: pathways to transformation. Nat. Rev. Cancer *10*, 550–560.

Munger, K., Howley, P., and DiMaio, D. (2007). Human Papillomavirus E6 and E7 Oncogenes. The Papillomaviruses 197–252.

Ohlenschläger, O., Seiboth, T., Zengerling, H., Briese, L., Marchanka, A., Ramachandran, R., Baum,

M., Korbas, M., Meyer-Klaucke, W., Dürst, M., *et al*. (2006). Solution structure of the partially folded high-risk human papilloma virus 45 oncoprotein E7. Oncogene *25*, 5953–5959.

Petoukhov, M.V., Franke, D., Shkumatov, A.V., Tria, G., Kikhney, A.G., Gajda, M., Gorba, C., Mertens, H.D.T., Konarev, P.V., and Svergun, D.I. (2012). New developments in the ATSAS program package for small-angle scattering data analysis. J. Appl. Crystallogr. *45*, 342–350.

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

Phelps, W.C., Yee, C.L., Munger, K., and Howley, P.M. (1988). The human papillomavirus type 16 E7 gene encodes transactivation and transformation functions similar to those of adenovirus E1A. Cell *53*, 539–547.

Roessle, M.W., Klaering, R., Ristau, U., Robrahn, B., Jahn, D., Gehrmann, T., Konarev, P., Round, A., Fiedler, S., Hermes, C., *et al*. (2007). Upgrade of the small-angle X-ray scattering beamline X33 at the European Molecular Biology Laboratory, Hamburg. J. Appl. Crystallogr. 190–194.

Round, A.R., Franke, D., Moritz, S., Huchler, R., Fritsche, M., Malthan, D., Klaering, R., Svergun, D.I., and Roessle, M. (2008). Automated sample-changing robot for solution scattering experiments at the EMBL Hamburg SAXS station X33. J. Appl. Crystallogr. *41*, 913–917.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soeding, J., *et al*. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology 539.

Svergun, D.I. (1992). Determination of the regularization parameter in indirect-transform methods using perceptual criteria. J. Appl. Crystallogr. *25*, 495–503.

Šali, A., and Blundell, T.L. (1993). Comparative Protein Modelling by Satisfaction of Spatial Restraints. J. Mol. Biol. *234*, 779–815.

Uversky, V.N., Roman, A., Oldfield, C.J., and Dunker, A.K. (2006). 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.

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

4.3.

The highly flexible and heterogeneous nature of HAdV E1A characterized at atomic resolution through NMR

Eduardo O. Calçada[a], *Isabella C. Felli*[a, b], *Tomáš Hošek*[a] *and Roberta Pierattelli*[a, b]

[a] Magnetic Resonance Center (CERM), University of Florence Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy

[b] Department of Chemistry "Ugo Schiff", University of Florence Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

To be submitted.

Mailing address
*CERM, University of Florence
Via Luigi Sacconi 6
50019 Sesto Fiorentino (FI), Italy
Email: felli@cerm.unifi.it
Email: pierattelli@cerm.unifi.it
Fax: +39 055 4574253

Abstract

The small DNA tumor viruses encode some of the most versatile hub proteins like the E1A protein from human Adenovirus (HAdV). The E1A protein is essential for productive viral infection in human cells and a vast amount of data are available on its interactions with host proteins. Up to now no high-resolution information on the full length E1A protein is available despite its important biological role.

Here we present the NMR characterization of the entire 289 residue long E1A protein from HAdV. The protein is very heterogeneous in terms of structural and dynamic properties with highly flexible modules. This study opens the way to characterize the manyinteractions in which this protein is involved.

Running title

NMR studies on HAdV E1A

Keywords

NMR; IDP; viral protein; adenovirus; SLiMs

Introduction

It is well known that the structural and dynamic properties of proteins are an essential aspect determining their function.[1-3] What is less known is all the modes through which highly disordered and flexible proteins or protein modules have an impact on protein function as, until recently, it was believed that only proteins characterized by well folded structures could be functional.[4;5] However, recent evidence accumulating at a very fast speed shows that also in the absence of a 3D structure proteins are functional and that actually a high extent of disorder and flexibility can provide functional advantages in a variety of different situations.

In particular the possibility to adopt different conformations may enable a protein to interact with many different partners.[6-8] This contributes to expanding our view on how proteins interact with each other based on many examples well documented in the PDB, such as with the presence of large complementary surfaces mediating protein-protein interactions. Of course more open, largely flexible conformations with large portions of exposed backbone may also be involved in interactions although in very different ways. Indeed, it has been suggested that short linear motifs, in other words short sequences of aminoacids with specific patterns, may be sufficient to mediate protein-protein interactions both in folded or in intrinsically disordered modules.[9] Of course, if short linear motives are sufficient to mediate protein-protein interactions, then the more exposed the backbone is the more short linear motives result exposed and available for mutual cross-talk between different proteins. Therefore the investigation of hub proteins and of how their function is modulated by their overall structural and dynamic properties and by short linear motives mediating interactions, is a theme under the spotlight of modern structural biology, with impressive examples appearing in the literature. [10;11]

In this frame, we would like here to focus on a viral protein, E1A from human adenovirus (HAdV). Viruses, with their small genomes, do need economic ways, in terms of genome usage, to interfere with host proteins. Therefore exploiting short linear motifs in intrinsically disordered protein modules does appear as an

appealing strategy towards this goal. HAdV is well characterized, routinely used in molecular biology techniques and a lot of information is available on it.[12-15] Among the few proteins encoded by HAdV we decided to focus on E1A, one of the first proteins produced and key for viral infection[16-18], because, despite all the information available on this virus and all the documented interactions of this protein with key players in host cells, there is no high-resolution structural and dynamic information available on the whole protein,[19] as the 3D structure of the protein could not be resolved by X-ray crystallography.

An interesting NMR investigation, instead, showed how a module of the protein is important in the interaction with a partner, CREB binding protein (CBP)[20], providing an interesting example of folding upon binding for a disordered protein. Therefore we decided to undertake the NMR investigation to the whole E1A in an attempt of understanding the molecular basis of how such a hub protein function.

Materials and methods

Cloning, expression and purification

The recombinant vector pET42-E1A containing the E1A gene from HAdV, was generously provided by Peter Pelka and Joe Mymryk, University of Western Ontario, London, Ontario, Canada.[21] The plasmid construct was amplified and the DNA screened showing the sequence of the E1A 13S from HAdV 2/5 hybrid. The final construct is basically the HAdV-2 13S E1A protein sequence where SER 276 is changed to PRO as in the HAdV-5 13S E1A protein sequence and GLU 148 is mutated to GLY.

The recombinant vector pET42-E1A was transformed into *E. coli* BL21(DE3)pLyS (Stratagene) for protein expression. One colony was selected to inoculate 50 ml LB medium containing ampicillin and chloramphenicol grown overnight at 37º C, 180 rpm. For ^{15}N and ^{13}C,^{15}N labelling 10 ml pre-inoculum were used to inoculate 1 l minimal 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 $MgSO_4$, 10.0 µM $ZnCl_2$, 1 mg/l each

of biotin and thiamin, 7.5 mM $(^{15}NH_4)_2SO_4$ and 11.1 mM glucose/$^{13}C_6$-glucose,) at 37ºC with constant agitation at 160 rpm. When optical density at 600nm reached 0.6, cells were induced with 0.5 mM final concentration of IPTG and allowed to grow for additional 16h at 25ºC. The cells were harvested by centrifugation at 8000g for 20 min and the cell pellet stored at -20ºC.

Frozen cells were thawed and suspended in 40 ml of equilibration buffer A (50mM Hepes, 300 mM NaCl, 5 mM imidazole, pH 8.0). Cells were disrupted by sonication on ice (at 40% sonication power) with cycles of 1 s with 10 s delay pulses for 30 min. Lysate cells were centrifuged at 30000 rpm for 30min at 4ºC and the supernatant was loaded on a 5ml Ni^{2+} HiTrap column (Amersham biosciences Europe) pre-equilibrated with buffer A under anaerobic conditions. Column was washed using 10x column volumes of buffer A with 50 mM imidazole and the E7 protein was eluted using buffer A with 250 mM imidazole. The eluted sample was then purified by size-exclusion chromatography using a HiLoad 16/60 Superdex 75 (Amersham biosciences Europe) column previously equilibrated with 10 mM HEPES buffer, 150 mM KCl, 10 µM $ZnCl_2$, 10 mM DTT. The column was refrigerated at 4 ^{0}C and sample was purified using a flow-rate of 1.5 ml mn^{-1}. The collected purified samples were checked both by SDS-PAGE and mass-spectrometry (MALDI and ESI) to confirm the molecular mass and purity E1A protein. All samples for NMR experiments were prepared in buffer containing 10 mM HEPES, 150 mM KCl, 10 µM $ZnCl_2$, 10 mM DTT. The E1A sample concentrations for NMR were 250µM, estimated using the molar extinction coefficient (7450 $M^{-1}cm^{-1}$) calculated from ExPASy ProtParam tool [http://web.expasy.org/ protparam/].[22] The sample pH was 7.5. 10% D_2O was added for the lock signal.

NMR Experiments

The ^{1}H detected NMR experiments for sequence specific assignment were acquired at 278 K on the 21.1 T Bruker AVANCE 900 NMR spectrometer, operating at 902.60 MHz for ^{1}H, equipped with cryogenically cooled TCI probe. A dataset of ^{1}H detected NMR experiments (^{1}H-^{15}N BEST-TROSY[23;24], BEST-TROSY

HNCO[25;26], BEST-TROSY HNCACO[25;26], BEST-TROSY HNCOCACB[25;26], BEST-TROSY HNCACB[25;26], BEST-TROSY HN(CA)NNH[25;27;28] and TROSY HN(COCA)NNH[29;30]) was necessary for sequence specific assignment of H^N, N, C', C^α, $C^ß$ resonances.

To characterize the structural and dynamic features of HAdV 2/5 E1A the following ^1H detected NMR experiments were acquired (at 278 K): ^{15}N relaxation experiments (R_1, R_2 and ^1H-^{15}N NOEs) on a 16.44 T Bruker AVANCE 700 spectrometer equipped with a cryogenically cooled probe.

The ^{13}C detected CON-IPAP experiment was acquired on the 16.44 T Bruker AVANCE 700 spectrometer equipped with a cryogenically cooled probe optimized for ^{13}C direct detection.

All experimental parameters used to acquire NMR experiments described above are reported in more detail in the Supplementary Material (Table S1).

Data were processed with TopSpin 2.0 and were analyzed with program ccpNMR[31]. The secondary structure propensity from chemical shifts was determined using the SSP program.[32] The ^{15}N relaxation rates (R_1 and R_2) were determined by fitting the cross-peak volumes measured as a function of variable delay, to single-exponential decay. ^1H-^{15}N NOE values were obtained as a ratio between peak volumes in spectra recorded with and without ^1H saturation.

Results

The primary sequence of E1A can be used to predict, through the available bioinformatics tools, the tendency of the different parts of the polypeptide chain to adopt more ordered or disordered conformations. As an example, predictions performed with the IUPred[33], Anchor[34] and PONDR-FIT[35] are reported in the Supplementary Material (Figure S1). These clearly indicate that E1A is characterized by significant structural heterogeneity, with largely disordered parts as well as largely ordered ones.

Aminoacid sequence alignment of E1A from different serotypes reveals four main conserved regions (CR1-CR4), as schematically indicated in Figure 1, which

are separated by less conserved regions.[36] Not only the CRs but also other parts of the E1A protein are involved in important protein-protein interactions with over 50 cellular factors through short linear motifs (SLiMs).[10;19] As an example, the SLiMs identified in the conserved regions are indicated in black in Figure 1. The high resolution structural and dynamic characterization of E1A represents thus an important step forward in the understanding of the molecular basis of its functional properties by providing experimental data to be compared with predictions and with a wealth of complementary information available on the protein through different molecular biology and biophysical techniques.

The 2D ^{1}H-^{15}N correlation experiment acquired on E1A is shown in Figure 2. The low chemical shift dispersion of the cross peaks observed confirms that the protein is characterized by a high extent of disorder and flexibility, as inferred from bioinformatic predictions (Figure S1).

The sequence specific assignment of the resonances represents thus a challenging task, both due to the extensive overlap observed as well as from the predictions that indicate a structurally heterogeneous system, characterized by highly disordered parts as well as by more structured modules. Moreover, the tendency of the protein to aggregate upon increasing concentration also poses limits in terms of protein concentrations that can be used for NMR experiments, restricting the choice among the available strategies (the 2D HN and 2D CON spectra are reported in Figure S2). With this in mind a series of 3D experiments based on amide proton detection were selected for the sequence specific assignment which included, in addition to the widely used ones (HNCO/HNCACO, CBCANH, CBCACONH) also two additional experiments which provide sequential correlations exploiting backbone nitrogen atoms and carbonyl carbons (HNCANNH, HNCOCANNH) which are known to provide valuable information to resolve ambiguities in particular when investigating highly disordered protein segments. The BEST-TROSY versions of the experiments were used.

While the extensive overlap observed in the 2D HN correlation experiments does not allow us to have an estimate on the number of identified correlations, the

analysis of the HNCO, thanks to the increased resolution provided by the carbonyl dimension, enables us to count the number of cross peaks that can be detected. In the present case 200 isolated cross peaks could be identified in the HNCO, which is significantly less than what expected on the basis of the primary sequence but still represents more than 80% of the number of residues that can be studies, significantly extending the atomic resolution characterization of E1A toward the entire polypeptide chain.

The careful analysis of the 3D spectra allowed us to achieve essentially complete sequence specific assignment of the polypeptide chain comprising CR1 and CR2 as well as the final module comprising CR4. Instead, our experiments are blind toward the CR3 region. The assigned residues include most of the residues involved in interactions with protein partners, as identified by ANCHOR, and the vast majority of the residues which are part of the SLiMs, indicating that the sequence specific assignment is going to open new avenues for the high resolution characterization of the many interactions in which E1A is involved.

The sequence specific assignment provides the first information to characterize the structural and dynamic properties of the protein through the analysis of backbone heteronuclear chemical shifts and of those of C^{β} carbons. Indeed by comparing experimental chemical shifts with those predicted assuming a completely random coil conformation it is possible to identify if secondary structural elements are partially populated. The resulting CSI values for C^{α}, C^{β}, C´ and N as well as the SSP score are reported in Figure 3.

Further observables providing useful information to characterize the structural and dynamic properties of the protein are ^{15}N relaxation data, in particular ^{15}N R_1, ^{15}N R_2 and ^{1}H-^{15}N NOE, reported in Figure 4. These show average values of 1.79 Hz for R_1 and of 4.89 Hz for R_2 (excluding in this case peak values in regions 15-23, 42-48, 70-75, 125-137, 265-270), which indicate that the protein is largely disordered. In addition the ^{1}H-^{15}N NOE values, all significantly lower than 1, also confirm that the protein is largely disordered. However, inspection of all the data does indicate that the protein is characterized by a significant structural

and dynamic heterogeneity, with several regions characterized by a propensity to populate more structured and/or less flexible conformations.

The first striking evidence of a region characterized by a more structured conformation comes from the analysis of chemical shifts. Indeed both CSI as well as SSP do indicate that in the initial, N-terminal part of the protein aminoacids 13-23 have a significant a-helical secondary structural propensity. This is confirmed by reduced local mobility in this region as identified through ^{15}N relaxation measurements, with R_2 values higher than the average (15-17, 19-20, 22-23) as well as with larger ^{1}H-^{15}N NOEs (14, 16-20, 22, 24). The rest of the polypeptide chain instead shows smaller CSI and SSP values. However inspection of the ^{15}N relaxation data does show several regions characterized by R_2 and NOE values larger than average which in general are indicative of less flexible parts of the polypeptide chain (40-44, 68-75, 123-137, 253-258, 261-270). Interestingly these regions are mainly found in the conserved regions. Starting from CR1, two SLiMs can be identified in this region (Figure 1) and are characterized by the presence of several hydrophobic residues which result characterized by lower local flexibility.

The interaction of this initial part of the polypeptide chain with the TAZ2 module of the CBP has been investigated by analyzing several short constructs and revealed that different regions are involved in binding to the partner with different properties.[20] These regions include the first part of the polypeptide where a high helical propensity has been identified (13-23) as well as two other regions comprising residues 59-65 and 72-75 that adopt an alpha helical conformation upon binding to the partner. Our data show that in the isolated full length E1A these fragments are largely disordered with a very weak tendency to adopt a a-helical conformation identified for aminoacids 68-70, right in between.

The other region characterized by ^{15}N relaxation parameters indicative of a less flexible and thus more structured part of the polypeptide chain are identified in CR2 (residues 123-137) and in CR4 (residues 253-258, 261-270). The LXCXE short linear motif (122-126), located in CR2, is is known to interact with the retinoblastoma binding protein (pRb), deregulating it.[10] Indeed, the interaction

of a short peptide mimicking this part of the polypeptide chain from the homologous protein E7 from the human papilloma virus with pRb has been characterized and showed that the LXCXE motif adopts an elongated conformation upon binding to pRb.[37] Interestingly the investigation of full length E7 showed a mild propensity of this part of the backbone as well as of neighboring ones, to assume an elongated conformation. In the present case, the aminoacid composition is slightly different, with a THR and a HIS and probably justifies why a local propensity, if any, for a helix turn is identified for this short fragment. Finally, the inspection of the data for CR4 also enables us to identify a very brief fragment characterized by a mild alpha helical secondary structural propensity for residues (268-270).

Conclusions

In summary, the high resolution information on the more flexible and disordered parts of E1A achieved here provides the basis for the high resolution characterization of all the interactions in which this protein is known to be involved. Thanks to vast amount of data available on the functional properties of this protein, of its interactions with many partners involved in key processes in the cell as well as to the lack, as of today, of high resolution information on the structural and dynamic properties of the protein, we hope to contribute to the understanding of how this protein interacts with many partners and more in general to unraveling new ways of protein-protein interactions enabling them to function as molecular hubs and of interfering with so many processes.

Acknowledgments

This work has been supported in part by the Joint Research Activity and Access to Research Infrastructures (BioNMR, contract 261863) and by the Marie Curie ITN programs (IDPbyNMR, contract 264257) in the EC 7th Framework.

Supplementary Material

Acquisition parameters of the NMR experiments; comparison of 2D HN and CON maps; bioinformatics prediction on the extent of disorder.

References

[1.] G. Wagner, S. Hyberts, J. W. Peng, in *NMR of Proteins*, Macmillan Press, New York **1993**, p. pp. 220-257.

[2.] G. Wagner, *Progr.NMR Spectrosc.* **1990,** *22* 101-139.

[3.] A. Mittermaier, L. E. Kay, *Science* **2006,** *312* 224-228.

[4.] P. E. Wright, H. J. Dyson, *Journal of Molecular Biology* **1999,** *293* 321-331.

[5.] B. He, K. Wang, Y. Liu, B. Xue, V. Uversky, A. K. Dunker, *Cell Res.* **2009,** *19* 929-949.

[6.] V. N. Uversky, C. J. Oldfield, A. K. Dunker, *J.Mol.Recognit.* **2005,** *18* 343-384.

[7.] M. Fuxreiter, I. Simon, P. Friedrich, P. Tompa, *J.Mol.Biol.* **2004,** *338* 1015-1026.

[8.] V. Uversky, C. J. Oldfield, A. K. Dunker, *Annu.Rev.Biophys.* **2008,** *37* 215-246.

[9.] C. M. Gould, F. Diella, A. Via, P. Puntervoll, C. Gemünd, S. Chabanis-Davidson, S. Michael, A. Sayadi, J. C. Bryne, C. Chica, M. Seiler, N. E. Davey, N. Haslam, R. J. Weatheritt, A. Budd, T. Huges, J. Pas, L. Rychlewski, G. Travé, R. Aasland, M. Helmer-Citterich, R. Linding, T. J. Gibson, *Nucleic Acids Res.* **2009,** *38* D167-D180.

[10.] N. E. Davey, G. Trave, T. J. Gibson, *Trends Biochem.Sci.* **2011,** *36* 159-169.

[11.] M. Wells, H. Tidow, T. J. Rutherford, P. Markwick, M. R. Jensen, E. Mylonas, D. I. Svergun, M. Blackledge, A. R. Fersht, *Proc.Natl.Acad.Sci.USA* **2008,** *105* 5762-5767.

[12.] P. H. Gallimore, *J.Gen.Virol* **1972,** *16* 99-102.

[13.] P. L. Stewart, R. M. Brunett, M. Cyrklaff, S. D. Fuller, *Cell* **1991,** *67* 145-154.

[14.] J. K. Mackey, P. M. Rigden, M. Green, *Proc.Natl Acad.Sci., USA* **1976,** *73* 4657-4661.

[15.] J. K. Mackey, W. S. M. Wold, P. Rigden, M. Green, *J.Virol* **1979,** *29* 1056-1064.

[16.] N. Jones, T. Shenk, *Proc.Natl Acad.Sci., USA* **1979,** *76* 3665-3669.

[17.] J. R. Nevins, H. S. Ginsberg, J. M. Blanchard, M. C. Wilson, J. E. Jr. Darnell, *J.Virol* **1979,** *32* 727-733.

[18.] P. Hearing, T. Shenk, *Cell* **1983,** *33* 695-703.

[19.] P. Pelka, J. N. G. Ablack, G. J. Fonseca, A. F. Yousef, J. S. Mymryk, *J.Virol.* **2008,** *82* 7252-7263.

[20.] J. C. Ferreon, M. A. Martinez-Yamout, H. J. Dyson, P. E. Wright, *Proc.Natl Acad.Sci., USA* **2009,** *106* 13260-13265.

[21.] 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.

[22.] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel, A. Bairoch, in *The Proteomics Protocols Handbook*Ed.: J. M. Walker), Humana Press, **2005**, p. pp. 571-607.

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

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

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

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

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

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

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

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

[31.] 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.

[32.] J. A. Marsh, V. K. Singh, Z. Jia, J. D. Forman-Kay, *Protein Sci.* **2006,** *15* 2795-2804.

[33.] Z. Dosztanyi, V. Csizmok, P. Tompa, I. Simon, *Bioinformatics* **2005,** *21* 3433-3434.

[34.] Z. Dosztanyi, B. Meszaros, I. Simon, *Bioinformatics* **2009,** *25* 2745-2746.

[35.] B. Xue, R. L. Jr. Dunbrack, R. W. Williams, A. K. Dunker, V. Uversky, *Biochim Biophys Acta* **2010,** *1804* 996-1010.

[36.] D. Kimelman, J. S. Miller, D. Porter, B. E. Roberts, *J.Virol.* **1985,** *53* 399-409.

[37.] J. O. Lee, A. A. Russo, N. P. Pavletich, *Nature* **1998,** *391* 859-865.

[38.] M. Kjaergaard, S. Brander, F. M. Poulsen, *J.Biomol.NMR* **2011,** *49* 139-149.

Figures

Figure 1. Schematic representation of HAdV 2/5 E1A protein amino acid sequence, and division into 4 conserved region (CR) with short linear motifs (SLiMs) indicated in black.[19]

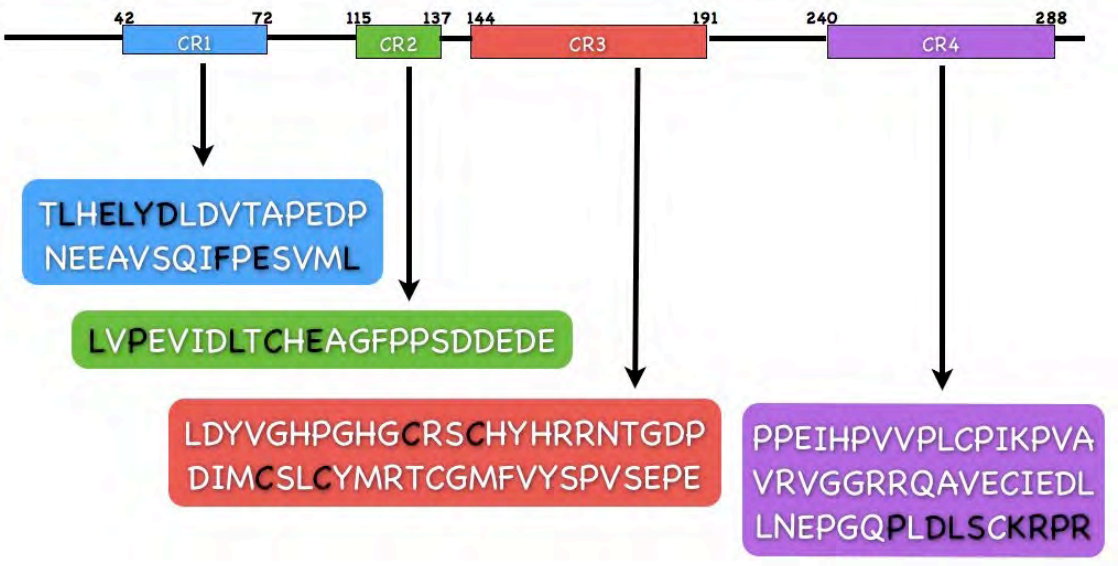


Figure 2. The 2D ^{1}H-^{15}N BEST-TROSY spectra recorded for HAdV 2/5 E1A. Experiment was recorded at 21.1 T and 278 K for a 0.2 mM E1A sample in 10 mM HEPES at pH 7.5, 150 mM KCl, 10 µM $ZnCl_2$, and 10 mM DTT. The complete chemical shift table is reported in Table S1.

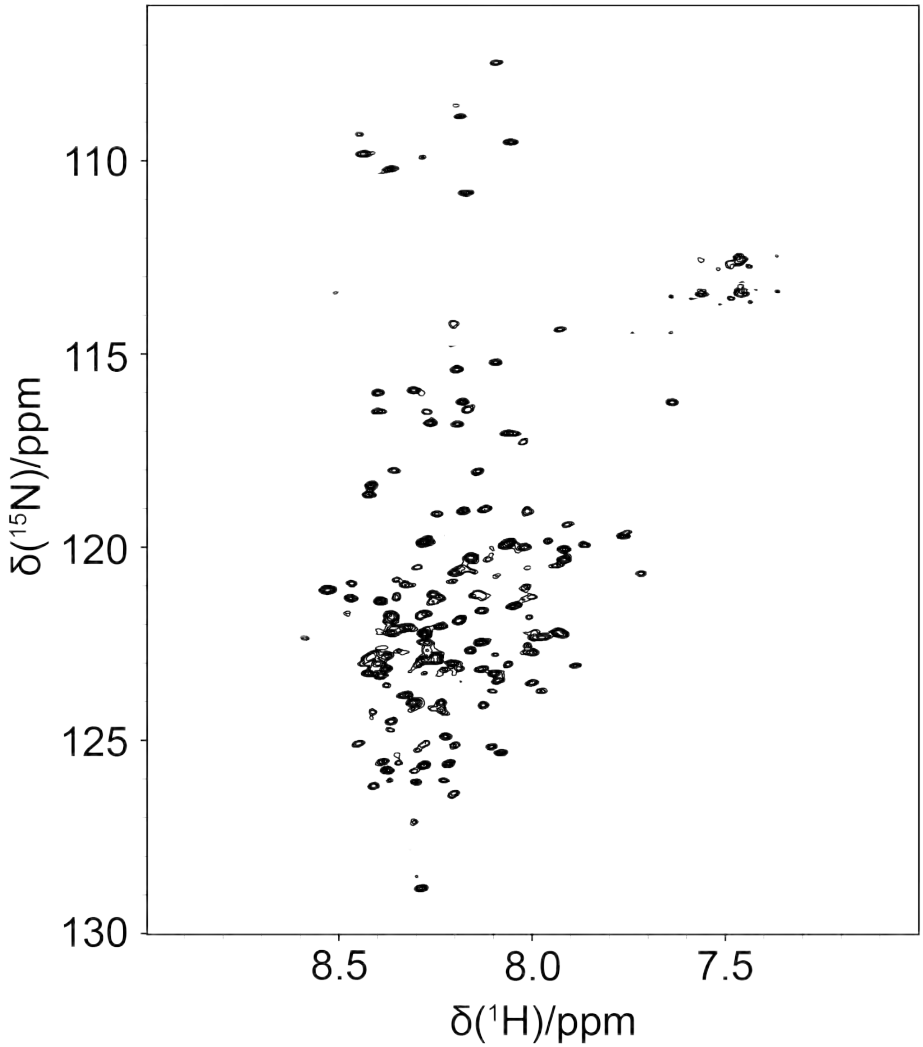
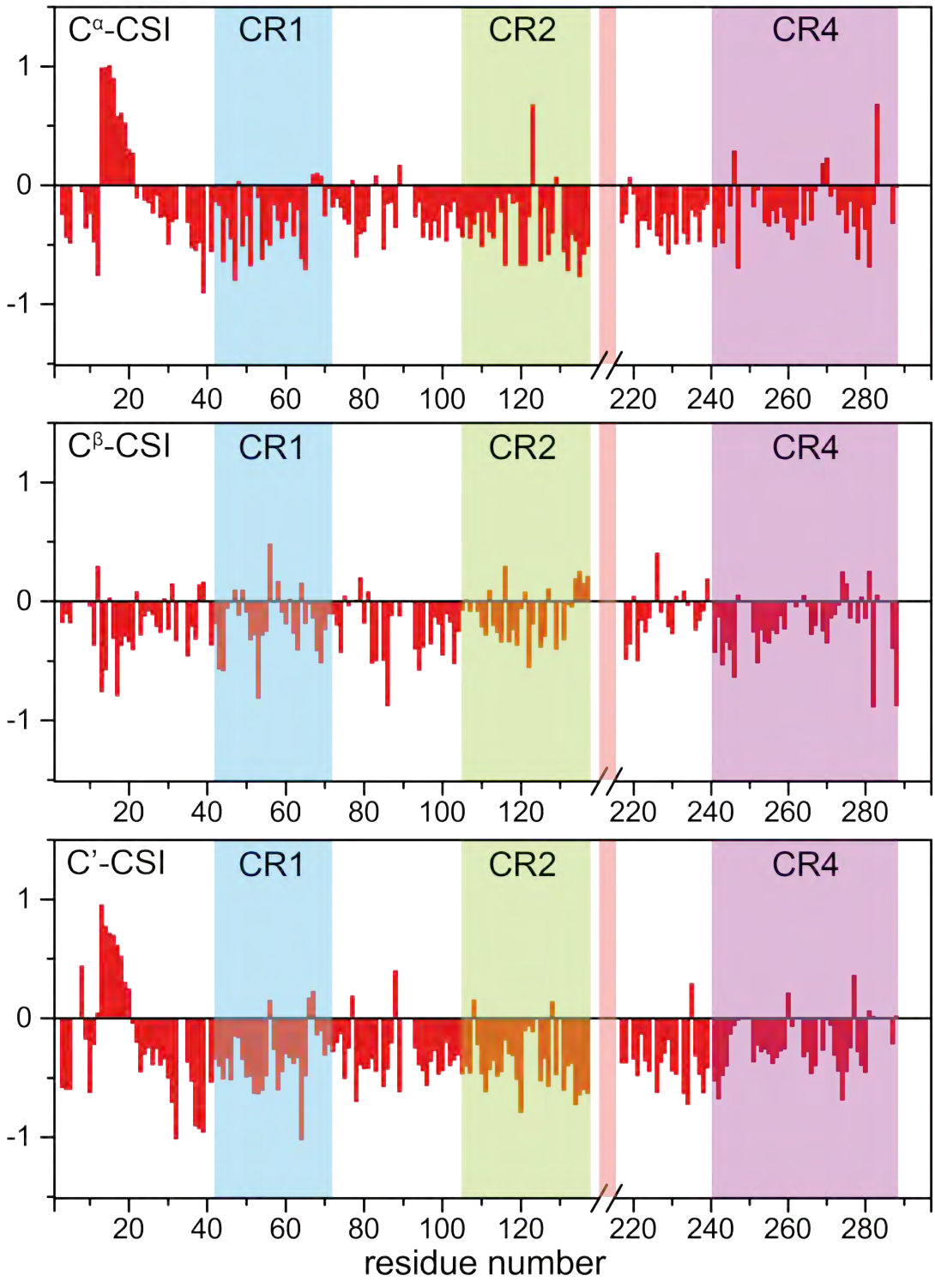
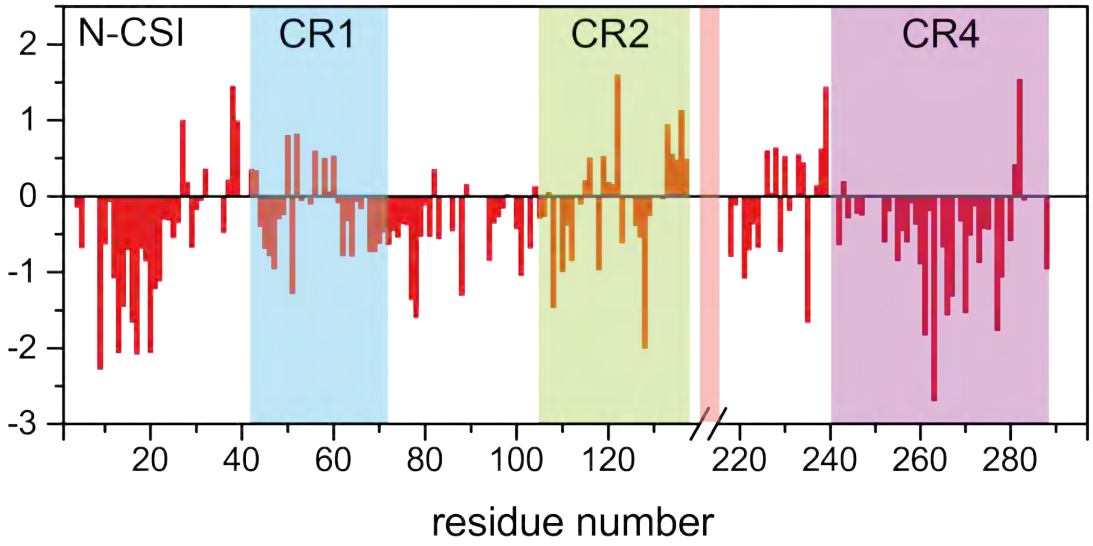


Figure 3. A) The chemical shift differences between the assigned resonances of HAdV 2/5 E1A with respect to random coil values corrected for the sequence, pH and temperature[38] for N^H, C', C^α and C^β nuclei. B) The Secondary structure propensity (SSP) score[32] obtained by combining C^α and C^β chemical shifts for the resonances of HAdV 2/5 E1A.

A)




N-CSI
CR1
CR2
CR4
2
1
0
-1
-2
-3
20
40
60
80
100
120
220
240
260
280
residue number

B)

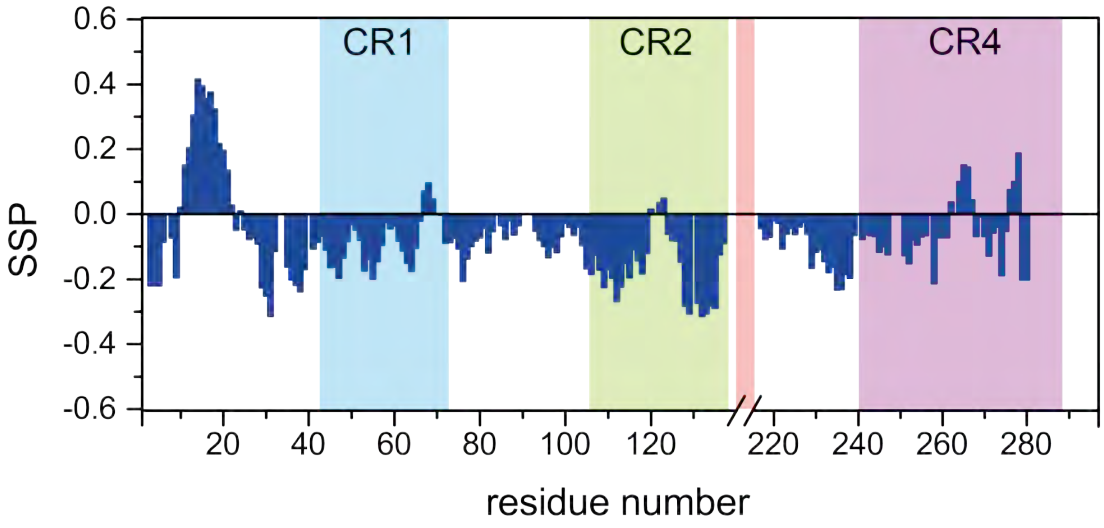
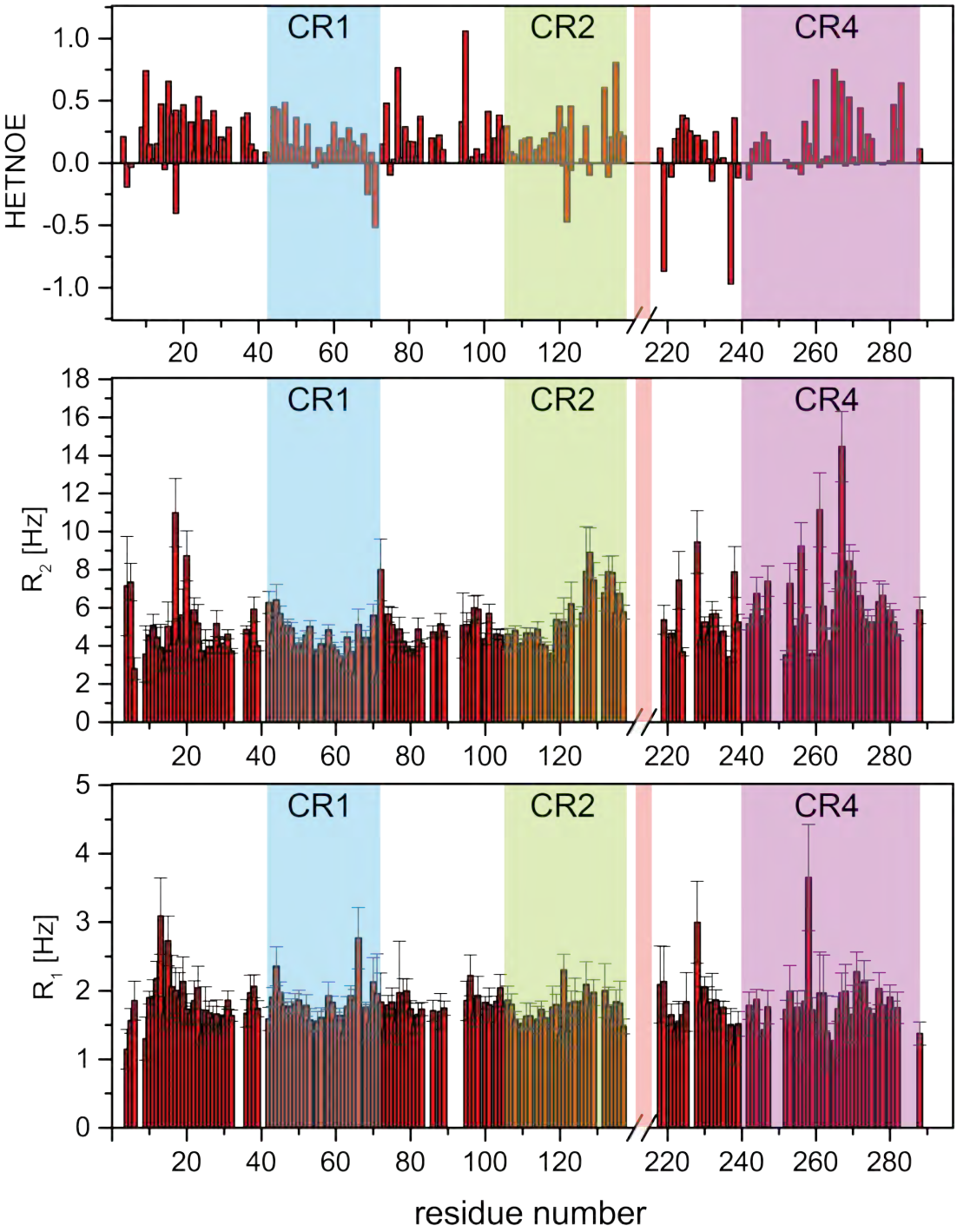

CR1
CR2
CR4
SSP
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
20
40
60
80
100
120
220
240
260
280
residue number

Figure 4. Relaxation data for backbone amide ^{15}N nuclei of HAdV 2/5 E1A as function of the residue number. From top to bottom: ^{1}H-^{15}N NOE values (HETNOE), ^{15}N transverse relaxation rates (R_2), and ^{15}N longitudinal relaxation rates (R_1). Experiments were recorded at 16.44 T and 278 K for a 0.1 mM E1A sample in 10 mM HEPES at pH 7.5, 150 mM KCl, 10 µM $ZnCl_2$, and 100 mM DTT.



Supplementary Material

The highly flexible and heterogeneous nature of HAdV E1A characterized at atomic resolution through 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 dynamic characterization of HAdV 2/5 E1A at 278 K.

The following parameters were used in all experiments. The ^1H carrier was placed at 4.7 ppm, the ^{15}N carrier at 118.5 ppm. Band selective ^{13}C pulses were given at 173 and 39 ppm to excite or invert C' and $C^{\alpha/ß}$ 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 experiments for sequence specific assignment were acquired with a Bruker AVANCE 900 spectrometer equipped with an inverse triple-resonance cryogenically cooled probehead (^1H-^{15}N BEST-TROSY[1;2], BEST-TROSY HNCO[3;4], BEST-TROSY HNCACO[3;4], BEST-TROSY HNCOCACB[3;4], BEST-TROSY HNCACB[3;4], BEST-TROSY HN(CA)NNH[3;5;6] and TROSY HN(COCA)NNH[7;8]). The other experiments (^{15}N R_1, ^{15}N R_2 and ^1H-^{15}N NOEs) were acquired with a Bruker AVANCE 700 spectrometer equipped with an inverse triple-resonance cryogenically cooled probehead.

Experiments	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]	512* (^{15}N)	1024 (^{13}C)		38	30		48	2.2
HCBCACON-IPAP[c]	128* (^{13}C)	64 (^{15}N)	1024 (^{13}C)	60	30	49	32	1
^1H detected								
^1H-^{15}N HSQC	256 (^{15}N)	1024 (^1H)		23	12		4	1
^1H-^{15}N BEST-TROSY	512 (^{15}N)	2048 (^1H)		30	12		8	0.2
BEST-TROSY HNCO	96 (^{13}C)	196 (^{15}N)	2048 (^1H)	8	24	15	8	0.2
BEST-TROSY HN(CA)CO	164 (^{13}C)	196 (^{15}N)	2048 (^1H)	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

Structure and dynamics

^{15}N T_1[d]	512 (^{15}N)	2048 (^{1}H)	24	16.1		8	3.0
^{15}N T_2[d]	512 (^{15}N)	2048 (^{1}H)	24	16.1		8	3.0
Steady-state heteronuclear ^{1}H-^{15}N-NOEs[d]	1024 (^{15}N)	2048 (^{1}H)	24	16.1		32	3.0

[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 850 ms. For the determination of R_2, 9 experiments were acquired changing the variable delay from 18 ms to 250 ms.
All 3D and 2D spectra were processed using the standard Bruker software (TopSpin 2.0) and analyzed through the ccpNMR (Collaborative Computing Project for NMR)[9]

Figure S1. ^{1}H-^{15}N HSQC and ^{13}C-^{15}N CON-IPAP spectra recorded for HAdV 2/5 E1A. Experiments were recorded at 16.4 T and 278 K for a 0.2 mM E1A sample in 10 mM HEPES buffer at pH 7.5, 150 mM KCl, 10 µM ZnCl2, and 10 mM DTT.

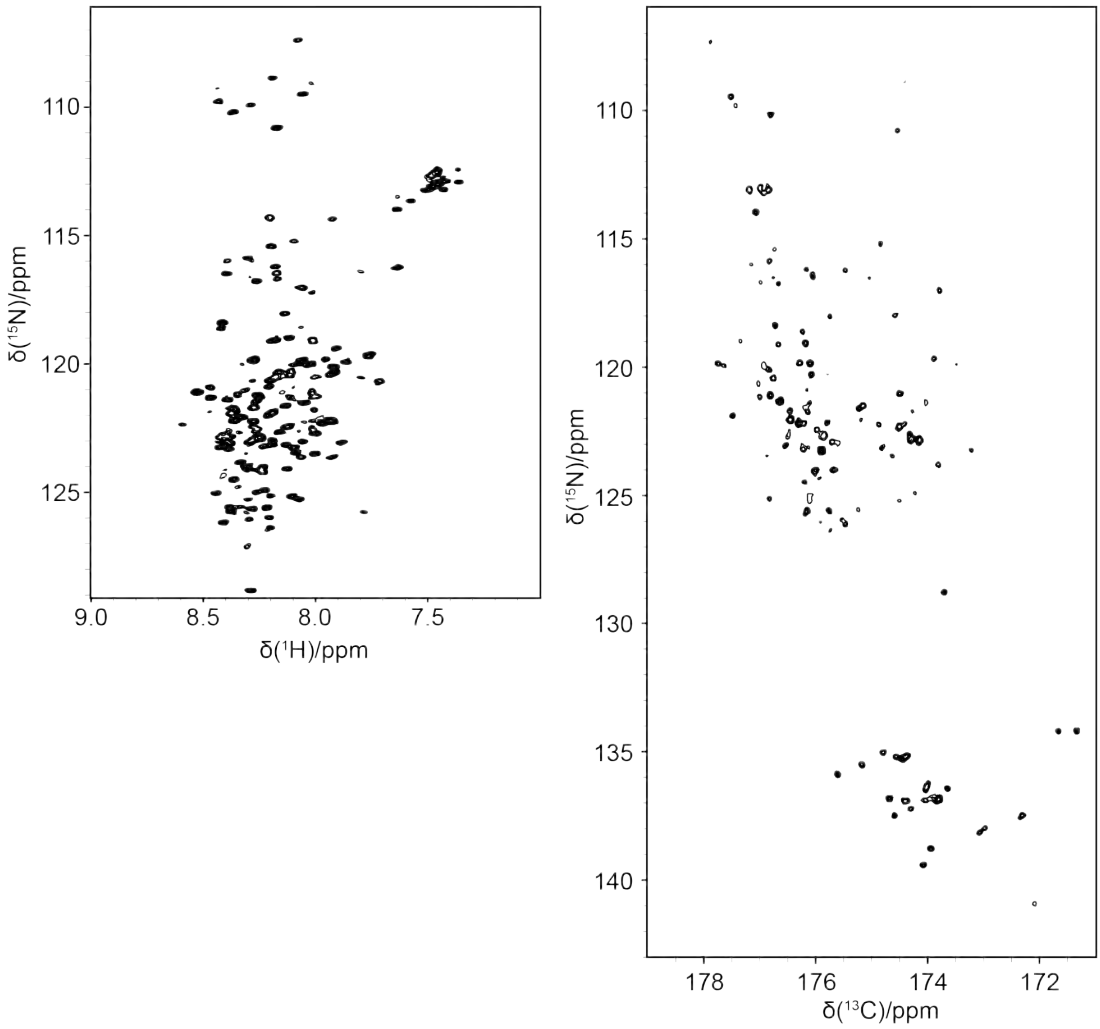
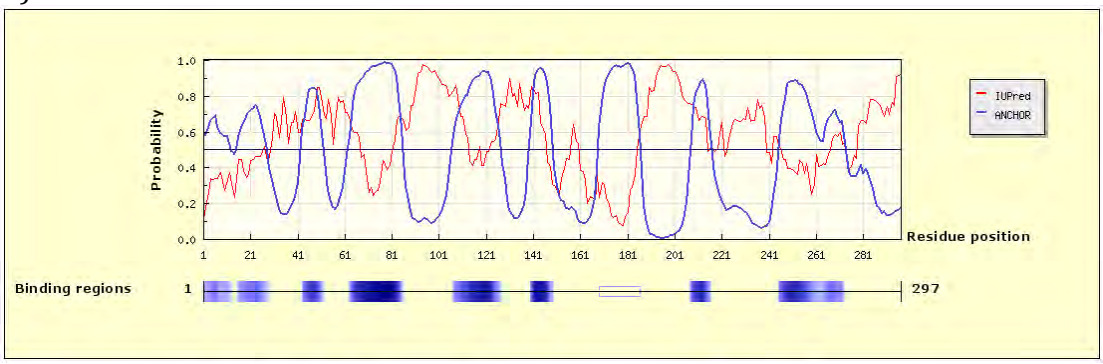
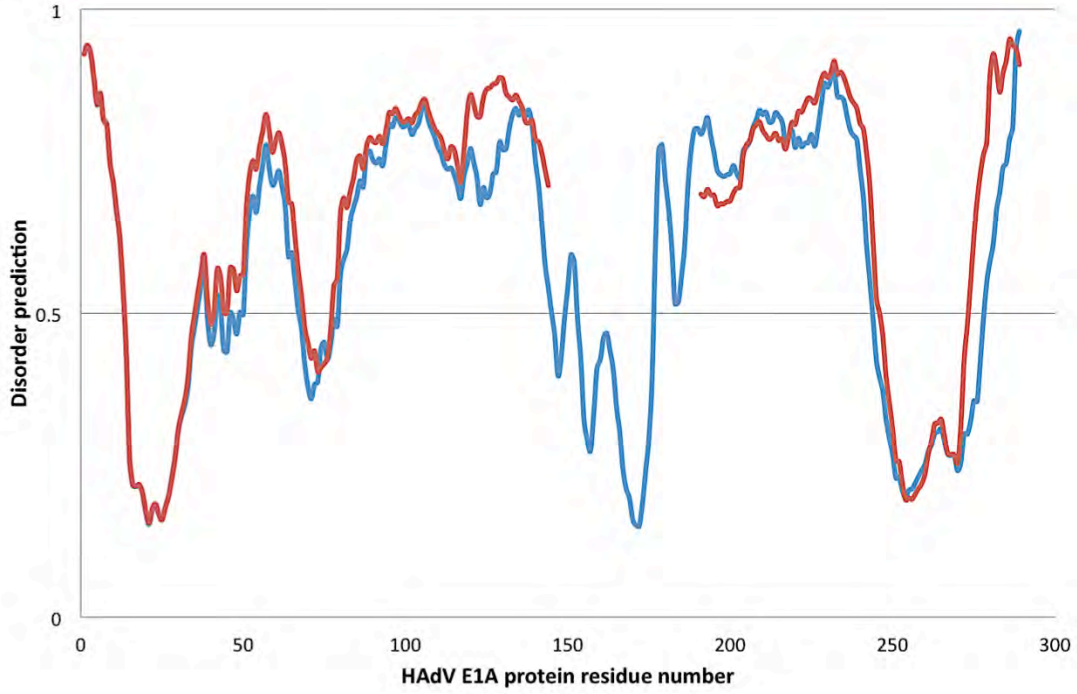


Figure S2. A) The prediction was done with the IUPred[10] and combined with the prediction of binding regions of E1A by ANCHOR[11].
B) The prediction was done with the metapredictor PONDR-FIT[12] comparing both sequences of the HAdV 12S and 13S protein sequences. The 12S E1A sequence doesn't contain the CR3 region, which is predicted to be folded. The alignment between both constructs was done taking in consideration the 46 residues of the CR3 region between residues 144 and 191 of the 13S E1A containing 289 residues.

A)



B)



Reference List

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

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

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

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

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

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

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

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

[9.] 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.

[10.] Z. Dosztanyi, V. Csizmok, P. Tompa, I. Simon, *Bioinformatics* **2005,** *21* 3433-3434.

[11.] Z. Dosztanyi, B. Meszaros, I. Simon, *Bioinformatics* **2009,** *25* 2745-2746.

[12.] B. Xue, R. L. Jr. Dunbrack, R. W. Williams, A. K. Dunker, V. Uversky, *Biochim Biophys Acta* **2010,** *1804* 996-1010.

[13.] J. L. Markley, A. Bax, Y. Arata, C. W. Hilbers, R. Kaptein, B. D. Sykes, P. E. Wright, K. Wüthrich, *J.Biomol.NMR* **1998,** *12* 1-23.

5. General conclusions and Perspectives

In this research project two key proteins from two different viral systems have been investigated. The Human Papilloma Virus E7 and the Human Adenovirus E1A proteins have been studied for the first time by NMR spectroscopy. These studies have been complemented by the results of other biophysical techniques. New insights for both proteins have been obtained, laying the basis for future investigations, in particular towards the understanding of the mode of interaction of these proteins with their targets.

5.1. Human Papilloma Virus E7 protein

The HPV 16 E7 protein has been studied for more than 30 years. Many attempts where done in the past to obtain high quality samples for NMR investigation of E7 from this particularly malignant phenotype without success. In this work the sample preparation was a crucial step. Many aspects were taken into consideration for getting high quality samples with various enrichment schemes of the full construct as well as for obtaining the folded E7 CR3 region, providing a body of data helpful to better understand the biophysics of this systems.

The full length HPV 16 E7 protein was analyzed by NMR spectroscopy for the first time. The dynamic properties of the entire protein were analyzed, revealing heterogeneous structural and dynamic properties well beyond the accepted schematic view of CR1 and CR2 being fully disordered and CR3 well structured. In addition, the NMR characterization provided an important platform for the study of the post-translational modifications and interactions of E7 from HPV 16 with the many partners known to interact with the different parts of the protein.

These data were completed by SAXS experiments using the full construct of E7 protein, which enabled to formulate an hypothesis on the mode the protein assembles in solution, eventually giving oligomeric forms, possibly explaining the different behavior of this protein with respect to the variants isolated from other viral types.

The current study opens the way to the atomic resolution characterization of the many interactions in which E7 is involved that may provide important information to understand the features that promote progression of cells to malignant and possibly , contribute to the design of new drugs.

5.2. Human Adenovirus E1A protein

The adenovirus E1A protein has been one of the most used models to study viral infections which promote cancer development. The lack of high-resolution information on the protein together with its relevance as a crucial modulator of host transcription upon infection prompted the study of this protein by NMR. E1A shares some important structural and dynamical similarities with E7, as revealed by NMR spectroscopy. The high resolution information obtained confirmed the presence of flexible and disordered parts, rich in short linear motives able to interact with the many targets of interaction. These comprise the CR1, CR2 and CR4 regions, while the CR3 region is expected to be characterized by an ordered structure.
As an immediate perspective of this work, we started the expression and the characterization of pRB and CBP/p300 proteins, which were shown to interact both with E7 and E1A. Preliminary data to initiate the expression of pRB have been obtained while for CBP the preparation of the protein has been already initiated. As the latter is 2442 amino acid residues, we choose one of the five intrinsically disordered regions, present among the structural domains that compose the architecture of CBP, the CBP-ID4. This domain is located right after a structured domain that was shown to interact with a short construct of E1A.
Concluding, the high resolution characterization of viral IDPs involved in oncogenic processes has been started with this research work. Thanks to vast amount of data available on the functional properties of the two key proteins selected, the high resolution information on their structural and dynamic properties achieved in this work contributes to the understanding of how they interact with many partners and more in general to unraveling new ways of protein-protein interactions, enabling them to function as molecular hubs and of interfering with many cellular processes.

6. References

Ablack, J.N.G., Pelka, P., Yousef, A.F., Turnell, A.S., Grand, R.J.A., and Mymryk, J.S. (2010). Comparison of E1A CR3-dependent transcriptional activation across six different human adenovirus subgroups. J. Virol. *84*, 12771–12781.

Alzari, P.M., Berglund, H., Berrow, N.S., Blagova, E., Busso, D., Cambillau, C., Campanacci, V., Christodoulou, E., Eiler, S., Fogg, M.J., *et al*. (2006). Implementation of semi-automated cloning and prokaryotic expression screening: the impact of SPINE. Acta Crystallogr. D Biol. Crystallogr. *62*, 1103–1113.

Arnau, J., Arnau, J., Lauritzen, C., Lauritzen, C., Petersen, G.E., Petersen, G.E., Pedersen, J., and Pedersen, J. (2006). Current strategies for the use of affinity tags and tag removal for the purification of recombinant proteins. Protein Expr. Purif. *48*, 1–13.

Babich, H., and Stotzky, G. (1978). Toxicity of zinc to fungi, bacteria, and coliphages: influence of chloride ions. Appl. Environ. Microbiol. *36*, 906–914.

Bandara, L.R., and La Thangue, N.B. (1991). Adenovirus E1a prevents the retinoblastoma gene product from complexing with a cellular transcription factor. Nature *351*, 494–497.

Banks, L., Pim, D., and Thomas, M. (2012). Human tumour viruses and the deregulation of cell polarity in cancer. Nat. Rev. Cancer *12*, 877–886.

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

Berezutskaya, E., Yu, B., Morozov, A., Raychaudhuri, P., and Bagchi, S. (1997). Differential regulation of the pocket domains of the retinoblastoma family proteins by the HPV16 E7 oncoprotein. Cell Growth Differ. *8*, 1277–1286.

Berk, A.J., Lee, F., Harrison, T., Williams, J., and Sharp, P.A. (1979). Pre-early adenovirus 5 gene product regulates synthesis of early viral messenger RNAs. Cell *17*, 935–944.

Bermel, W., Bertini, I., Csizmok, V., Felli, I.C., Pierattelli, R., and Tompa, P. (2009). H-start for exclusively heteronuclear NMR spectroscopy: the case of intrinsically disordered proteins. J. Magn. Reson. *198*, 275–281.

Bermel, W., Bertini, I., Felli, I.C., Gonnelli, L., Koźmiński, W., Piai, A., Pierattelli, R., and Stanek, J. (2012). Speeding up sequence specific assignment of IDPs. J. Biomol. NMR *53*, 293–301.

Bermel, W., Bertini, I., Felli, I.C., Kümmerle, R., and Pierattelli, R. (2006a). Novel 13C direct detection experiments, including extension to the third dimension, to perform the complete assignment of proteins. J. Magn. Reson. *178*, 56–64.

Bermel, W., Bertini, I., Felli, I.C., Lee, Y.-M., Luchinat, C., and Pierattelli, R. (2006b). Protonless NMR experiments for sequence-specific assignment of backbone nuclei in unfolded proteins. J. Am. Chem. Soc. *128*, 3918–3919.

Bernadó, P., and Svergun, D.I. (2012a). Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. Mol. BioSyst. *8*, 151–167.

Bernadó, P., and Svergun, D.I. (2012b). Analysis of intrinsically disordered proteins by small-angle X-ray scattering. Methods Mol. Biol. *896*, 107–122.

Bernadó, P., Mylonas, E., Petoukhov, M.V., Blackledge, M., and Svergun, D.I. (2007a). Structural characterization of flexible proteins using small-angle X-ray scattering. J. Am. Chem. Soc. *129*, 5656–5664.

Bernadó, P., Mylonas, E., Petoukhov, M.V., Blackledge, M., and Svergun, D.I. (2007b). Structural characterization of flexible proteins using small-angle X-ray scattering. J. Am. Chem. Soc. *129*, 5656–5664.

Bernard, H.-U., Burk, R.D., Chen, Z., van Doorslaer, K., Hausen, H.Z., and de Villiers, E.-M. (2010). Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. Virology *401*, 70–79.

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

Bertini, I., McGreevy, K.S., and Parigi, G. (2012). NMR of Biomolecules (John Wiley & Sons).

Block, H., Maertens, B., Spriestersbach, A., Brinker, N., Kubicek, J., Fabis, R., Labahn, J., and Schäfer, F. (2009). Immobilized-metal affinity chromatography (IMAC): a review. Meth. Enzymol. *463*, 439–473.

Boulet, G., Horvath, C., Broeck, D.V., and Sahebali, S. (2007). Human papillomavirus: E6 and E7 oncogenes. Int. J. Biochem. Cell Biol. *39*, 2006–2011.

Braun, D., Wider, G., and Wuethrich, K. (1994). Sequence-Corrected 15N "Random Coil" Chemical Shifts. J. Am. Chem. Soc. *116*, 8466–8469.

Brito, R.M.M., and Vaz, W.L.C. (1986). Determination of the Critical Micelle Concentration of Surfactants Using the Fluorescent-Probe N-Phenyl-1-Naphthylamine. Anal. Biochem. *152*, 250–255.

Bruton, R.K., Pelka, P., Mapp, K.L., Fonseca, G.J., Torchia, J., Turnell, A.S., Mymryk, J.S., and Grand, R.J.A. (2008). Identification of a second CtBP binding site in adenovirus type 5 E1A conserved region 3. J. Virol. *82*, 8476–8486.

Buchan, D.W.A., Minneci, F., Nugent, T.C.O., Bryson, K., and Jones, D.T. (2013). Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Research *41*, gkt381–W357.

Burgess, R., Richard, R., and Murray, P. (2009). Refolding Solubilized Inclusion Body Proteins. Methods Enzymol 259–282.

Calçada, E.O., Felli, I.C., Hošek, T., and Pierattelli, R. (2013). The heterogeneous structural behavior of E7 from HPV16 revealed by NMR spectroscopy. ChemBioChem *14*, 1876–1882.

Chemes, L.B., Glavina, J., Faivovich, J., de Prat-Gay, G., and Sánchez, I.E. (2012). Evolution of Linear Motifs within the Papillomavirus E7 Oncoprotein. J. Mol. Biol. *422*, 336–346.

Chemes, L.B., Sánchez, I.E., Smal, C., and de Prat-Gay, G. (2010). 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.

Cheng, J., Saigo, H., and Baldi, P. (2006). Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching. Proteins: Structure, Function, and Bioinformatics *62*, 617–629.

Chinami, M., Sasaki, S., and Hachiya, N. (1994). Functional oligomerization of purified human papillomavirus types 16 and 6b E7 proteins expressed in Escherichia coli.

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

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

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

Cohen, M.J., Yousef, A.F., Massimi, P., Fonseca, G.J., Todorovic, B., Pelka, P., Turnell, A.S., Banks, L., and Mymryk, J.S. (2013). Dissection of the C-terminal region of E1A redefines the roles of CtBP and other cellular targets in oncogenic transformation. J. Virol. *87*, 10348–10355.

Conte, Lo, M., and Carroll, K.S. (2013). The redox biochemistry of protein sulfenylation and sulfinylation. Journal of Biological Chemistry *288*, 26480–26488.

Coutard, B., Danchin, E.G.J., Oubelaid, R., Canard, B., and Bignon, C. (2012). Single pH buffer refolding screen for protein from inclusion bodies. Protein Expr. Purif. *82*, 352–359.

Cowieson, N.P., Wensley, B., Listwan, P., Hume, D.A., Kobe, B., and Martin, J.L. (2006). An automatable screen for the rapid identification of proteins amenable to refolding. Proteomics *6*, 1750–1757.

Culp, J.S., and Webster, L.C. (1988). 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.

Das, U., Hariprasad, G., Ethayathulla, A.S., Manral, P., Das, T.K., Pasha, S., Mann, A., Ganguli, M., Verma, A.K., Bhat, R., *et al*. (2007). Inhibition of Protein Aggregation: Supramolecular Assemblies of Arginine Hold the Key. PLoS ONE *2*.

Davey, N.E., Haslam, N.J., Shields, D.C., and Edwards, R.J. (2010). SLiMFinder: a web server to find novel, significantly over-represented, short protein motifs. Nucleic Acids Research *38*, W534–W539.

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

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

Di Marco, V.B., and Bombi, G.G. (2006). Electrospray mass spectrometry (ESI-MS) in the study of metal–ligand solution equilibria. Mass Spectrometry Reviews *25*, 347–379.

Diaz, A.A., Tomba, E., Lennarson, R., Richard, R., Bagajewicz, M.J., and Harrison, R.G. (2010). Prediction of protein solubility in Escherichia coli using logistic regression. Biotechnology and Bioengineering *105*, 374–383.

Diella, F., Haslam, N., Chica, C., Budd, A., Michael, S., Brown, N.P., Trave, G., and Gibson, T.J. (2008). Understanding eukaryotic linear motifs and their role in cell signaling and regulation. Front Biosci *13*, 6580–6603.

Dimmock, N., Easton, A., and Leppard, K. (2007). Introduction to Modern Virology (Wiley-Blackwell).

Dinkel, H., Chica, C., Via, A., Gould, C.M., Jensen, L.J., Gibson, T.J., and Diella, F. (2011). Phospho.ELM: a database of phosphorylation sites—update 2011. Nucleic Acids Research *39*,

D261–D267.

Dosztányi, Z., and Tompa, P. (2008). Prediction of Protein Disorder. Structural Proteomics *426*, 103–115.

Dosztányi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics *21*, 3433–3434.

Dunker, A.K., and Obradovic, Z. (2001). The protein trinity - linking function and disorder. Nat. Biotechnol. *19*, 805–806.

Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., Hipps, K.W., *et al.* (2001). Intrinsically disordered protein. J. Mol. Graph. Model. *19*, 26–59.

Dunker, A.K., Babu, M.M., Barbar, E., Blackledge, M., Bondos, S.E., Dosztányi, Z., Dyson, H.J., Forman-Kay, J.D., Fuxreiter, M., Gsponer, J., *et al.* (2013). What's in a name? Why these proteins are intrinsically disordered. Intrinsically Disordered Proteins *1*, e24157.

Dunker, A.K., Oldfield, C.J., Meng, J., Romero, P., Yang, J.Y., Chen, J.W., Vacic, V., Obradovic, Z., and Uversky, V.N. (2008a). The unfoldomics decade: an update on intrinsically disordered proteins. BMC Genomics *9 Suppl 2*, S1.

Dunker, A.K., Silman, I., Uversky, V.N., and Sussman, J.L. (2008b). Function and structure of inherently disordered proteins. Curr. Opin. Struct. Biol. *18*, 756–764.

Dyson, H.J., and Wright, P.E. (2001). Nuclear magnetic resonance methods for elucidation of structure and dynamics in disordered states. Meth. Enzymol. *339*, 258–270.

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

Dyson, N., Guida, P., McCall, C., and Harlow, E. (1992a). Adenovirus-E1a Makes 2 Distinct Contacts with the Retinoblastoma Protein. The Journal of Virology *66*, 4606–4611.

Dyson, N., Guida, P., Münger, K., and Harlow, E. (1992b). Homologous Sequences in Adenovirus E1a and Human Papillomavirus E7 Proteins Mediate Interaction with the Same Set of Cellular Proteins. The Journal of Virology *66*, 6893–6902.

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

El-Aneed, A., Cohen, A., and Banoub, J. (2009). Mass spectrometry, review of the basics: Electrospray, MALDI, and commonly used mass analyzers. Applied Spectroscopy Reviews.

Emanuelsson, O., Brunak, S.O.R., Heijne, von, G., and Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc *2*, 953–971.

Esposito, D., and Chatterjee, D.K. (2006). Enhancement of soluble protein expression through the use of fusion tags. Current Opinion in Biotechnology *17*, 353–358.

Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.-Y., Pieper, U., and Šali, A. (2006). Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics *UNIT 5.6*, Unit5.6.

Felli, I.C., and Pierattelli, R. (2012). Recent progress in NMR spectroscopy: Toward the study of intrinsically disordered proteins of increasing size and complexity. Iubmb Life *64*, 473–481.

Fera, D., and Marmorstein, R. (2012). Different Regions of the HPV-E7 and Ad-E1A Viral Oncoproteins Bind Competitively but through Distinct Mechanisms to the CH1 Transactivation Domain of p300. Biochemistry *51*, 9524–9534.

Ferguson, B., Krippl, B., Andrisani, O., Jones, N., Westphal, H., and Rosenberg, M. (1985). E1A 13S and 12S mRNA products made in Escherichia coli both function as nucleus-localized transcription activators but do not directly bind DNA. Mol. Cell. Biol. *5*, 2653–2661.

Ferreon, A.C.M., Ferreon, J.C., Wright, P.E., and Deniz, A.A. (2013). Modulation of allostery by protein intrinsic disorder. Nature *498*, 390–394.

Ferreon, J.C., Martinez-Yamout, M.A., Dyson, H.J., and Wright, P.E. (2009). Structural basis for subversion of cellular control mechanisms by the adenoviral E1A oncoprotein. Proc. Natl. Acad. Sci. U.S.a. *106*, 13260–13265.

Ferrè, F., and Clote, P. (2006). DiANNA 1.1: an extension of the DiANNA web server for ternary cysteine classification. Nucleic Acids Research *34*, W182–W185.

Ferron, F., Longhi, S., Canard, B., and Karlin, D. (2006). A practical overview of protein disorder prediction methods. Proteins *65*, 1–14.

Franke, D., and Svergun, D.I. (2009). DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. J. Appl. Crystallogr. *42*, 342–346.

Frisch, S.M., and Mymryk, J.S. (2002). Adenovirus-5 E1A: paradox and paradigm. Nat. Rev. Mol. Cell Biol. *3*, 441–452.

Fuchs, M., Gerber, J., Drapkin, R., Sif, S., Ikura, T., Ogryzko, V., Lane, W.S., Nakatani, Y., and Livingston, D.M. (2001). The p400 Complex Is an Essential E1A Transformation Target. Cell *106*, 297–307.

Fuxreiter, M., Tompa, P., and Simon, I. (2007). Local structural disorder imparts plasticity on linear motifs. Bioinformatics *23*, 950–956.

Fuxreiter, M., Tompa, P., Simon, I.A.N., Uversky, V.N., Hansen, J.C., and Asturias, F.J. (2008). Malleable machines take shape in eukaryotic transcriptional regulation. Nat. Chem. Biol. *4*, 728–737.

Gallimore, P.H. (1972). Tumour Production in Immunosuppressed Rats with Cells Transformed in vitro by Adenovirus Type 2. Journal of General Virology *16*, 99–102.

Garcia-Alai, M.M., Alonso, L.G., and de Prat-Gay, G. (2007). The N-terminal module of HPV16 E7 is an intrinsically disordered domain that confers conformational and recognition plasticity to the oncoprotein. Biochemistry *46*, 10405–10412.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., and Bairoch, A. (2005). Protein Identification and Analysis Tools on the ExPASy Server (Totowa, NJ: Humana Press).

Getz, E.B., Xiao, M., Chakrabarty, T., Cooke, R., Cooke, R., Selvin, P.R., and Selvin, P.R. (1999). A Comparison between the Sulfhydryl Reductants Tris(2-carboxyethyl)phosphine and Dithiothreitol for Use in Protein Biochemistry. Anal. Biochem. *273*, 73–80.

Grzesiek, S., and Bax, A. (1993). Measurement of amide proton exchange rates and NOEs with water in 13C/15N-enriched calcineurin B. J. Biomol. NMR *3*, 627–638.

Guinier, A. (1939). La diffraction des rayons X aux tres petits angles: applications a l'etude de phenomenes ultramicroscopiques. Annales De Physique, EDP Sciences.

Gupta, B.B. (1983). Determination of native and denatured milk proteins by high-performance size exclusion chromatography. Journal of Chromatography A *282*, 463–475.

Halldén, G. (2012). Adenoviral Gene Expression and Replication in Human Tumor Explant Models. Oncolytic Viruses *797*, 65–78.

Harpaz, Y., Gerstein, M., and Chothia, C. (1994). Volume changes on protein folding. Structure *2*, 641–649.

Hateboer, G., Gennissen, A., Ramos, Y., Kerkhoven, R.M., Sonntagbuck, V., Stunnenberg, H.G., and Bernards, R. (1995). Bs69, a Novel Adenovirus E1a-Associated Protein That Inhibits E1a Transactivation. Embo J. *14*, 3159–3169.

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

Haynes, C., Oldfield, C.J., Ji, F., Klitgord, N., Cusick, M.E., Radivojac, P., Uversky, V.N., Vidal, M., and Iakoucheva, L.M. (2006). Intrinsic Disorder Is a Common Feature of Hub Proteins from Four Eukaryotic Interactomes. PLoS Comput. Biol. *2*, e100.

He, B., Wang, K., Liu, Y., Xue, B., Uversky, V.N., and Dunker, A.K. (2009). Predicting intrinsic disorder in proteins: an overview. Cell Res. *19*, 929–949.

Hearing, P., and Shenk, T. (1983). The adenovirus type 5 E1A transcriptional control region contains a duplicated enhancer element. Cell *33*, 695–703.

Huh, K., Zhou, X., Hayakawa, H., Cho, J.-Y., Libermann, T.A., Jin, J., Harper, J.W., and Münger, K. (2007). Human papillomavirus type 16 E7 oncoprotein associates with the cullin 2 ubiquitin ligase complex, which contributes to degradation of the retinoblastoma tumor suppressor. The Journal of Virology *81*, 9737–9747.

Hulo, C., de Castro, E., Masson, P., Bougueleret, L., Bairoch, A., Xenarios, I., and Le Mercier, P. (2011). ViralZone: a knowledge resource to understand virus diversity. Nucleic Acids Research *39*, D576–D582.

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

Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z., and Dunker, A.K. (2004). The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Research *32*, 1037–1049.

Ignatova, Z., and Gierasch, L.M. (2006). Inhibition of protein aggregation in vitro and in vivo by a natural osmoprotectant. Proceedings of the National Academy of Sciences *103*, 13357–13361.

Ishida, T., and Kinoshita, K. (2008). Prediction of disordered regions in proteins based on the meta approach. Bioinformatics *24*, 1344–1348.

Jacques, D.A., Guss, J.M., Svergun, D.I., and Trewhella, J. (2012). Publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution. Acta Crystallogr. D Biol. Crystallogr. *68*, 620–626.

Jana, S., and Deb, J.K. (2005). Strategies for efficient production of heterologous proteins in Escherichia coli. Appl Microbiol Biotechnol *67*, 289–298.

Jones, N., and Shenk, T. (1979a). An adenovirus type 5 early gene function regulates expression of other early viral genes. Proc. Natl. Acad. Sci. U.S.a. *76*, 3665–3669.

Jones, N., and Shenk, T. (1979b). Isolation of adenovirus type 5 host range deletion mutants defective for transformation of rat embryo cells. Cell *17*, 683–689.

Jumpertz, T., Tschapek, B., Infed, N., Smits, S.H.J., Ernst, R., and Schmitt, L. (2011). High-throughput evaluation of the critical micelle concentration of detergents. Anal. Biochem. *408*, 64–70.

Katzen, F. (2007). Gateway ®recombinational cloning: a biological operating system. Expert Opin. Drug Discov. *2*, 571–589.

Kay, L.E., Torchia, D.A., and Bax, A. (1989). Backbone dynamics of proteins as studied by 15N inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. Biochemistry *28*, 8972–8979.

Kelly, S.M., Jess, T.J., and Price, N.C. (2005). How to study proteins by circular dichroism. Biochimica Et Biophysica Acta (BBA) - Proteins and Proteomics *1751*, 119–139.

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

Kindermann, B., Döring, F., Fuchs, D., Pfaffl, M.W., and Daniel, H. (2005). Effects of increased cellular zinc levels on gene and protein expression in HT-29 cells. Biometals *18*, 243–253.

Konarev, P.V., Petoukhov, M.V., Volkov, V.V., and Svergun, D.I. (2006). ATSAS 2.1, a program package for small-angle scattering data analysis. J. Appl. Crystallogr. *39*, 277–286.

Konarev, P.V., Volkov, V.V., Sokolova, A.V., Koch, M.H.J., and Svergun, D.I. (2003). PRIMUS: a Windows PC-based system for small-angle scattering data analysis. J. Appl. Crystallogr. *36*, 1277–1282.

Krezel, A., Latajka, R., Bujacz, G.D., and Bal, W. (2003). Coordination properties of tris(2-carboxyethyl)phosphine, a newly introduced thiol reductant, and its oxide. Inorg Chem *42*, 1994–2003.

Krogh, A., Larsson, B., Heijne, von, G., and Sonnhammer, E.L.L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. J. Mol. Biol. *305*, 567–580.

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

Lesley, S.A. (2009). Parallel Methods for Expression and Purification. In Methods in Enzymology, (Elsevier), pp. 767–785.

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

Lieutaud, P., Canard, B., and Longhi, S. (2008). MeDor: a metaserver for predicting protein disorder. BMC Genomics *9 Suppl 2*, S25.

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

Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., and Russell, R.B. (2003a). Protein Disorder Prediction. Structure *11*, 1453–1459.

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

Linn, S. (2009). Strategies and Considerations for Protein Purifications. In Methods in Enzymology, (Elsevier), pp. 9–19.

Liu, A., Riek, R., Wider, G., Schroetter, von, C., Zahn, R., and Wüthrich, K. (2000). NMR experiments for resonance assignments of 13C, 15N doubly-labeled flexible polypeptides: application to the human prion protein hPrP(23-230). J. Biomol. NMR *16*, 127–138.

Liu, F., and Green, M.R. (1994). Promoter targeting by adenovirus E1a through interaction with different cellular DNA-binding domains. Nature *368*, 520–525.

Liu, X., Clements, A., Zhao, K., and Marmorstein, R. (2006). Structure of the Human Papillomavirus E7 Oncoprotein and Its Mechanism for Inactivation of the Retinoblastoma Tumor Suppressor. Journal of Biological Chemistry *281*, 578–586.

Louis-Jeune, C., Andrade-Navarro, M.A., and Perez-Iratxeta, C. (2012). Prediction of protein secondary structure from circular dichroism using theoretically derived spectra. Proteins *80*, 374–381.

Magnan, C.N., Randall, A., and Baldi, P. (2009). SOLpro: accurate sequence-based prediction of protein solubility. Bioinformatics *25*, 2200–2207.

Malhotra, A. (2009). Tagging for Protein Expression. In Methods in Enzymology, (Elsevier), pp. 239–258.

Malloy, A.H., and Carson, E.C. (2008). Oncogene Proteins: New Research (Nova Science Publishers).

Marsh, J.A., Singh, V.K., Jia, Z., and Forman-Kay, J.D. (2006). Sensitivity of secondary structure propensities to sequence differences between α- and γ-synuclein: Implications for fibrillation. Protein Sci. *15*, 2795–2804.

McLaughlin-Drubin, M.E., and Münger, K. (2009). The human papillomavirus E7 oncoprotein. Virology *384*, 335–344.

McLaughlin-Drubin, M.E., Meyers, J., and Munger, K. (2012). Cancer associated human papillomaviruses. Curr Opin Virol *2*, 459–466.

Mertens, H.D.T., and Svergun, D.I. (2010). Structural characterization of proteins and complexes using small-angle X-ray solution scattering. Journal of Structural Biology *172*, 128–141.

Miller, M.S., Pelka, P., Fonseca, G.J., Cohen, M.J., Kelly, J.N., Barr, S.D., Grand, R.J.A., Turnell, A.S., Whyte, P., and Mymryk, J.S. (2012). Characterization of the 55-residue protein encoded by the 9S E1A mRNA of species C adenovirus. J. Virol. *86*, 4222–4233.

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

Mittag, T., and Forman-Kay, J.D. (2007). Atomic-level characterization of disordered protein ensembles. Curr. Opin. Struct. Biol. *17*, 3–14.

Mittermaier, A., and Kay, L.E. (2006). Review - New tools provide new insights in NMR studies of protein dynamics. Science *312*, 224–228.

Monastyrskyy, B., Fidelis, K., Moult, J., Tramontano, A., and Kryshtafovych, A. (2011). Evaluation of disorder predictions in CASP9. Proteins: Structure, Function, and Bioinformatics *79*, 107–118.

Moody, C.A., and Laimins, L.A. (2010a). Human papillomavirus oncoproteins: pathways to

transformation. Nat. Rev. Cancer *10*, 550–560.

Moody, C.A., and Laimins, L.A. (2010b). Human papillomavirus oncoproteins: pathways to transformation. Nat. Rev. Cancer *10*, 550–560.

Munger, K., Howley, P., and DiMaio, D. (2007). Human Papillomavirus E6 and E7 Oncogenes. The Papillomaviruses 197–252.

Münger, K., Phelps, W.C., Bubb, V., Howley, P.M., and Schlegel, R. (1989). The E6 and E7 genes of the human papillomavirus type 16 together are necessary and sufficient for transformation of primary human keratinocytes. J. Virol. *63*, 4417–4421.

Münger, K., Baldwin, A., Edwards, K.M., Hayakawa, H., Nguyen, C.L., Owens, M., Grace, M., and Huh, K. (2004). Mechanisms of human papillomavirus-induced oncogenesis. The Journal of Virology *78*, 11451–11460.

Narisawa Saito, M., and Kiyono, T. (2007). Basic mechanisms of high-risk human papillomavirus-induced carcinogenesis: Roles of E6 and E7 proteins. Cancer Science *98*, 1505–1511.

Nevins, J.R., Ginsberg, H.S., Blanchard, J.M., Wilson, M.C., and Darnell, J.E. (1979). Regulation of the primary expression of the early adenovirus transcription units. The Journal of Virology *32*, 727–733.

Oates, M.E., Romero, P., Ishida, T., Ghalwash, M., Mizianty, M.J., Xue, B., Dosztányi, Z., Uversky, V.N., Obradovic, Z., Kurgan, L., *et al.* (2013). D²P²: database of disordered protein predictions. Nucleic Acids Research *41*, D508–D516.

Ohlenschläger, O., Seiboth, T., Zengerling, H., Briese, L., Marchanka, A., Ramachandran, R., Baum, M., Korbas, M., Meyer-Klaucke, W., Dürst, M., *et al.* (2006). Solution structure of the partially folded high-risk human papilloma virus 45 oncoprotein E7. Oncogene *25*, 5953–5959.

Outten, C.E., and O'Halloran, A.T.V. (2001). Femtomolar Sensitivity of Metalloregulatory Proteins Controlling Zinc Homeostasis. Science *292*, 2488–2492.

Passerini, A., Lippi, M., and Frasconi, P. (2011). MetalDetector v2.0: predicting the geometry of metal binding sites from protein sequence. Nucleic Acids Research *39*, W288–W292.

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

Pelka, P., Ablack, J.N.G., Torchia, J., Turnell, A.S., Grand, R.J.A., and Mymryk, J.S. (2009). Transcriptional control by adenovirus E1A conserved region 3 via p300/CBP. Nucleic Acids Research *37*, 1095–1106.

Pelka, P., Miller, M.S., Cecchini, M., Yousef, A.F., Bowdish, D.M., Dick, F., Whyte, P., and Mymryk, J.S. (2011). Adenovirus E1A directly targets the E2F/DP-1 complex. J. Virol. *85*, 8841–8851.

Peng, J.W., and Wagner, G. (1994). Investigation of protein motions via relaxation measurements. Meth. Enzymol. *239*, 563–596.

Perricaudet, M., rvi, G.O.R.A.A., Virtanen, A., and Pettersson, U. (1979). Structure of two spliced mRNAs from the transforming region of human subgroup C adenoviruses. Nature *281*, 694–696.

Petoukhov, M.V., Franke, D., Shkumatov, A.V., Tria, G., Kikhney, A.G., Gajda, M., Gorba, C., Mertens, H.D.T., Konarev, P.V., and Svergun, D.I. (2012). New developments in the ATSAS

program package for small-angle scattering data analysis. J. Appl. Crystallogr. *45*, 342–350.

Petoukhov, M.V., and Svergun, D.I. (2013). Applications of small-angle X-ray scattering to biomacromolecular solutions. Int. J. Biochem. Cell Biol. *45*, 429–437.

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

Phelps, W.C., Yee, C.L., Munger, K., and Howley, P.M. (1988). The human papillomavirus type 16 E7 gene encodes transactivation and transformation functions similar to those of adenovirus E1A. Cell *53*, 539–547.

Prilusky, J., Felder, C.E., Zeev-Ben-Mordehai, T., Rydberg, E.H., Man, O., Beckmann, J.S., Silman, I., and Sussman, J.L. (2005). FoldIndex©: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics *21*, 3435–3438.

Qoronfleh, M.W., Hesterberg, L.K., and Seefeldt, M.B. (2007). Confronting high-throughput protein refolding using high pressure and solution screens. Protein Expr. Purif. *55*, 209–224.

Radivojac, P., Iakoucheva, L.M., Oldfield, C.J., Obradovic, Z., Uversky, V.N., and Dunker, A.K. (2007). Intrinsic disorder and functional proteomics. Biophys. J. *92*, 1439–1456.

Ratnayake, S., Dias, I.H.K., Lattman, E., and Griffiths, H.R. (2013). Stabilising cysteinyl thiol oxidation and nitrosation for proteomic analysis. J Proteomics *92*, 160–170.

Rehder, D.S., and Borges, C.R. (2010). Cysteine sulfenic acid as an intermediate in disulfide bond formation and nonenzymatic protein folding. Biochemistry *49*, 7748–7755.

Rezaei Ghaleh, N., Blackledge, M., and Zweckstetter, M. (2012). Intrinsically Disordered Proteins: From Sequence and Conformational Properties toward Drug Discovery. ChemBioChem *13*, 930–950.

Roessle, M.W., Klaering, R., Ristau, U., Robrahn, B., Jahn, D., Gehrmann, T., Konarev, P., Round, A., Fiedler, S., Hermes, C., *et al*. (2007). Upgrade of the small-angle X-ray scattering beamline X33 at the European Molecular Biology Laboratory, Hamburg. J. Appl. Crystallogr. 190–194.

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

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

Round, A.R., Franke, D., Moritz, S., Huchler, R., Fritsche, M., Malthan, D., Klaering, R., Svergun, D.I., and Roessle, M. (2008). Automated sample-changing robot for solution scattering experiments at the EMBL Hamburg SAXS station X33. J. Appl. Crystallogr. *41*, 913–917.

Schenk, P.M., Baumann, S., Mattes, R., and Steinbiss, H. (1995). Improved high-level expression system for eukaryotic genes in Escherichia coli using T7 RNA polymerase and rare ArgtRNAs. Biotechniques *19*, 196–200.

Schiffman, M., Castle, P., Jeronimo, J., and Rodriguez, A. (2007). Human papillomavirus and cervical cancer. The Lancet *370*, 890–907.

Schlessinger, A., Punta, M., Yachdav, G., Kajan, L., and Rost, B. (2009). Improved Disorder Prediction by Combination of Orthogonal Approaches. PLoS ONE *4*, e4433.

Schwarzinger, S., Kroon, G.J., Foss, T.R., Chung, J., Wright, P.E., and Dyson, H.J. (2001). Sequence-dependent correction of random coil NMR chemical shifts. J. Am. Chem. Soc. *123*,

2970–2978.

Shatzman, A.R. (1995). Expression systems.

Shetty, V., Spellman, D.S., and Neubert, T.A. (2007). Characterization by tandem mass spectrometry of stable cysteine sulfenic acid in a cysteine switch peptide of matrix metalloproteinases. J Am Soc Mass Spectrom *18*, 1544–1551.

Sickmeier, M., Hamilton, J.A., LeGall, T., Vacic, V., Cortese, M.S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V.N., *et al*. (2007). DisProt: the Database of Disordered Proteins. Nucleic Acids Research *35*, D786–D793.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soeding, J., *et al*. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology 539.

Singh, S.M., Singh, S.M., Panda, A.K., and Panda, A.K. (2005). Solubilization and refolding of bacterial inclusion body proteins. Journal of Bioscience and Bioengineering *99*, 303–310.

Smialowski, P., Martin-Galiano, A.J., Mikolajka, A., Girschick, T., Holak, T.A., and Frishman, D. (2007). Protein solubility: sequence based prediction and experimental verification. Bioinformatics *23*, 2536–2542.

Spera, S., and Bax, A. (1991). Empirical correlation between protein backbone conformation and C.alpha. and C.beta. 13C nuclear magnetic resonance chemical shifts. J. Am. Chem. Soc. *113*, 5490–5492.

Stephens, C., and Harlow, E. (1987). Differential splicing yields novel adenovirus 5 E1A mRNAs that encode 30 kd and 35 kd proteins. Embo J. *6*, 2027–2035.

Stewart, P.L., Burnett, R.M., Cyrklaff, M., and Fuller, S.D. (1991). Image reconstruction reveals the complex molecular organization of adenovirus. Cell *67*, 145–154.

Subramanian, T., Zhao, L., and Chinnadurai, G. (2013). Interaction of CtBP with adenovirus E1A suppresses immortalization of primary epithelial cells and enhances virus replication during productive infection. Virology 313–320.

Svergun, D.I. (1992). Determination of the regularization parameter in indirect-transform methods using perceptual criteria. J. Appl. Crystallogr. *25*, 495–503.

Šali, A., and Blundell, T.L. (1993). Comparative Protein Modelling by Satisfaction of Spatial Restraints. J. Mol. Biol. *234*, 779–815.

Testa, L., Brocca, S., Santambrogio, C., DUrzo, A., Habchi, J., Longhi, S., Uversky, V.N., and Grandori, R. (2013). Extracting structural information from charge-state distributions of intrinsically disordered proteins by non-denaturing electrospray-ionization mass spectrometry. Intrinsically Disordered Proteins *1*, e25068.

Theillet, F.-X., Kalmar, L., Tompa, P., Han, K.-H., Selenko, P., Dunker, A.K., Daughdrill, G.W., and Uversky, V.N. (2013). The alphabet of intrinsic disorder I. Act like a Pro: On the abundance and roles of proline residues in intrinsically disordered proteins. Intrinsically Disordered Proteins *1*, 0–12.

Todorovic, B., Hung, K., Massimi, P., Avvakumov, N., Dick, F.A., Shaw, G.S., Banks, L., and Mymryk, J.S. (2012). Conserved region 3 of human papillomavirus 16 E7 contributes to deregulation of the retinoblastoma tumor suppressor. J. Virol. *86*, 13313–13323.

Tompa, P. (2012). On the supertertiary structure of proteins. Nat. Chem. Biol. *8*, 597–600.

Tompa, P., and Fersht, A. (2009). Structure and Function of Intrinsically Disordered Proteins (Chapman and Hall/CRC).

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

Tong, K.I., Yamamoto, M., and Tanaka, T. (2008). A simple method for amino acid selective isotope labeling of recombinant proteins in E. coli. J. Biomol. NMR *42*, 59–67.

Uversky, V.N., and Longhi, S. (2011). Flexible Viruses: Structural Disorder in Viral Proteins.

Uversky, V.N. (2011). Intrinsically disordered proteins from A to Z. Int. J. Biochem. Cell Biol. *43*, 1090–1103.

Uversky, V.N. (2013a). A decade and a half of protein intrinsic disorder: biology still waits for physics. Protein Sci. *22*, 693–724.

Uversky, V.N. (2013b). Unusual biophysics of intrinsically disordered proteins. Biochimica Et Biophysica Acta (BBA) - Proteins and Proteomics *1834*, 932–951.

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

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

Uversky, V.N., Oldfield, C.J., and Dunker, A.K. (2008). Intrinsically disordered proteins in human diseases: introducing the D2 concept. Annu Rev Biophys *37*, 215–246.

Uversky, V.N., Roman, A., Oldfield, C.J., and Dunker, A.K. (2006). 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.

Varadi, M., Kosol, S., Lebrun, P., Valentini, E., Blackledge, M., Dunker, A.K., Felli, I.C., Forman-Kay, J.D., Kriwacki, R.W., Pierattelli, R., *et al*. (2013). pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. Nucleic Acids Research.

Vincentelli, R., Canaan, S., Campanacci, V., Valencia, C., Maurin, D., Frassinetti, F., Scappucini-Calvo, L., Bourne, Y., Cambillau, C., and Bignon, C. (2004). High-throughput automated refolding screening of inclusion bodies. Protein Science *13*, 2782–2792.

Wagner, G. (1993). NMR relaxation and protein mobility. Curr. Opin. Struct. Biol. *3*, 748–754.

Wang, F., Marshall, C.B., and Ikura, M. (2013). Transcriptional/epigenetic regulator CBP/p300 in tumorigenesis: structural and functional versatility in target recognition. Cell. Mol. Life Sci. 1–20.

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

Ward, J.J., McGuffin, L.J., Bryson, K., Buxton, B.F., and Jones, D.T. (2004b). The DISOPRED server for the prediction of protein disorder. Bioinformatics *20*, 2138–2139.

Wong, H.K., and Ziff, E.B. (1994). Complementary functions of E1a conserved region 1 cooperate with conserved region 3 to activate adenovirus serotype 5 early promoters. The Journal of Virology *68*, 4910–4920.

Wright, P.E., and Dyson, H.J. (1999). Intrinsically unstructured proteins: re-assessing the

protein structure-function paradigm. J. Mol. Biol. *293*, 321–331.

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

Xue, B., Ganti, K., Rabionet, A., Banks, L., and Uversky, V.N. (2013). Disordered Interactome of Human Papillomavirus. Curr. Pharm. Des. 1–19.

Zamanian, M., and La Thangue, N.B. (1992). Adenovirus E1a prevents the retinoblastoma gene product from repressing the activity of a cellular transcription factor. Embo J. *11*, 2603–2610.

Zerler, B., Roberts, R.J., Mathews, M.B., and Moran, E. (1987). Different Functional Domains of the Adenovirus E1a Gene Are Involved in Regulation of Host-Cell Cycle Products. Mol. Cell. Biol. *7*, 821–829.

Zhang, O., Forman-Kay, J.D., Shortle, D., and Kay, L.E. (1997). Triple-resonance NOESY-based experiments with improved spectral resolution: applications to structural characterization of unfolded, partially folded and folded proteins. J. Biomol. NMR *9*, 181–200.

7. Supplementary data

Table 6.1. Summary of proteins studied in this work. Legend: MM, molecular mass in Daltons; pI, isoelectric point; ε, extinction coefficient factor.

	HPV16 E7	HPV16 E7 CR3	HAdV 2/5 E1A13S	HAdV 2 E1A12S	CBP-ID4
MM (Da)	12087	6401	33176	26500	22539
pI	4.7	6.0	4.9	4.4	12.1
ε (M^{-1} cm^{-1})	5960	1490	7450	1490	6990
N°residues	106	58	299	243	211
Ala (A)	3 (3%)	2 (3%)	13 (4%)	13 (5%)	19 (9%)
Arg (R)	3 (3%)	3 (5%)	18 (6%)	14 (6%)	13 (6%)
Asn (N)	2 (2%)	1 (2%)	7 (2%)	6 (2%)	7 (3%)
Asp (D)	10 (9%)	4 (7%)	17 (6%)	14 (6%)	1 (1%)
Cys (C)	7 (7%)	6 (10%)	12 (4%)	7 (3%)	0 (0%)
Gln (Q)	5 (5%)	2 (3%)	8 (3%)	8 (3%)	34 (16%)
Glu (E)	10 (9%)	2 (3%)	33 (11%)	30 (12%)	3 (1%)
Gly (G)	5 (5%)	3 (5%)	16 (5%)	11 (5%)	11 (5%)
His (H)	10 (9%)	2 (3%)	19 (6%)	7 (3%)	3 (1%)
Ile (I)	5 (5%)	4 (7%)	12 (4%)	11 (5%)	4 (2%)
Leu (L)	12 (11%)	6 (10%)	24 (8%)	21 (9%)	7 (3%)
Lys (K)	2 (2%)	2 (3%)	3 (1%)	3 (1%)	1 (1%)
Met (M)	3 (3%)	1 (2%)	7 (2%)	4 (2%)	12 (6%)
Phe (F)	1 (1%)	2 (3%)	7 (2%)	5 (2%)	2 (1%)
Pro (P)	6 (6%)	3 (5%)	46 (16%)	43 (18%)	46 (22%)
Ser (S)	5 (5%)	4 (7%)	20 (7%)	18 (7%)	17 (8%)
Thr (T)	9 (9%)	6 (10%)	11 (4%)	9 (4%)	15 (7%)
Trp (W)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (1%)
Tyr (Y)	4 (4%)	1 (2%)	5 (2%)	1 (0%)	1 (1%)
Val (V)	4 (4%)	4 (7%)	21 (7%)	18 (7%)	14 (7%)

Sequence alignment of E1A 13S proteins between HAdV2 and HAdV5

```
HAdV2    E1A  MRHIICHGGVITEEMAASLLDQLIEEVLADNLPPPSHFEPPTLHELYDLDVTAPEDPNEE
HAdV2/5  E1A  MRHIICHGGVITEEMAASLLDQLIEEVLADNLPPPSHFEPPTLHELYDLDVTAPEDPNEE
HAdV5    E1A  MRHIICHGGVITEEMAASLLDQLIEEVLADNLPPPSHFEPPTLHELYDLDVTAPEDPNEE
              ************************************************************

HAdV2    E1A  AVSQIFPESVMLAVQEGIDLFTFPPAPGSPEPPHLSRQPEQPEQRALGPVSMPNLVPEVI
HAdV2/5  E1A  AVSQIFPESVMLAVQEGIDLFTFPPAPGSPEPPHLSRQPEQPEQRALGPVSMPNLVPEVI
HAdV5    E1A  AVSQIFPDSVMLAVQEGIDLLTFPPAPGSPEPPHLSRQPEQPEQRALGPVSMPNLVPEVI
              *******:***********:****************************************

HAdV2    E1A  DLTCHEAGFPPSDDEDEEGEEFVLDYVEHPGHGCRSCHYHRRNTGDPDIMCSLCYMRTCG
HAdV2/5  E1A  DLTCHEAGFPPSDDEDEEGEEFVLDYVGHPGHGCRSCHYHRRNTGDPDIMCSLCYMRTCG
HAdV5    E1A  DLTCHEAGFPPSDDEDEEGEEFVLDYVEHPGHGCRSCHYHRRNTGDPDIMCSLCYMRTCG
              ************************** ******************************

HAdV2    E1A  MFVYSPVSEPEPEPEPEPEPEPARPTRRPKLVPAILRRPTSPVSRECNSSTDSCDSGPSNTP
HAdV2/5  E1A  MFVYSPVSEPEPEPEPEPEPEPARPTRRPKLVPAILRRPTSPVSRECNSSTDSCDSGPSNTP
HAdV5    E1A  MFVYSPVSEPEPEPEPEPEPEPARPTRRPKMAPAILRRPTSPVSRECNSSTDSCDSGPSNTP
              ****************************:.******************************

HAdV2    E1A  PEIHPVVPLCPIKPVAVRVGGRRQAVECIEDLLNESGQPLDLSCKRPRP
HAdV2/5  E1A  PEIHPVVPLCPIKPVAVRVGGRRQAVECIEDLLNEPGQPLDLSCKRPRP
HAdV5    E1A  PEIHPVVPLCPIKPVAVRVGGRRQAVECIEDLLNEPGQPLDLSCKRPRP
              ********************************** *************
```

The sequence alignment shows the differences between the E1A 13S proteins from the HAdV2, the HAdV5 and HAdV 2/5 hybrid that we studied in the present work (summarized on table 6.2). The hybrid between both constructs is essentially the sequence of HAdV2 13S E1A protein with exception for the residue number 276 which is a proline originated from the HAdV5 13S E1A protein sequence. There is also a glycine residue on position 148 instead of the expected glutamic acid typical from both HAdV 2 and 5 types.

Table 6.2. Comparison of amino acid differences between E1A protein sequences from HAdV2, HAdV5 and the resulting HAdV2/5 hybrid

HAdV 13S E1A residue number	HAdV2 E1A	HAdV5 E1A	HAdV2/5 E1A hybrid
68	E	D	E
81	F	L	F
148	E	E	G
209	L	M	L
210	V	A	V
276	S	P	P

Sequence alignment of HAdV2 E1A 12S and 13S proteins

```
HAdV2 13S    MRHIICHGGVITEEMAASLLDQLIEEVLADNLPPPSHFEPPTLHELYDLDVTAPEDPNEE
HAdV2 12S    MRHIICHGGVITEEMAASLLDQLIEEVLADNLPPPSHFEPPTLHELYDLDVTAPEDPNEE
             ************************************************************

HAdV2 13S    AVSQIFPESVMLAVQEGIDLFTFPPAPGSPEPPHLSRQPEQPEQRALGPVSMPNLVPEVI
HAdV2 12S    AVSQIFPESVMLAVQEGIDLFTFPPAPGSPEPPHLSRQPEQPEQRALGPVSMPNLVPEVI
             ************************************************************

HAdV2 13S    DLTCHEAGFPPSDDEDEEGEEFVLDYVGHPGHGCRSCHYHRRNTGDPDIMCSLCYMRTCG
HAdV2 12S    DLTCHEAGFPPSDDEDEEG-----------------------------------------
             ******************

HAdV2 13S    MFVYSPVSEPEPEPEPEPEPARPTRRPKLVPAILRRPTSPVSRECNSSTDSCDSGPSNTP
HAdV2 12S    -----PVSEPEPEPEPEPEPARPTRRPKLVPAILRRPTSPVSRECNSSTDSCDSGPSNTP
                  ***************************************************

HAdV2 13S    PEIHPVVPLCPIKPVAVRVGGRRQAVECIEDLLNESGQPLDLSCKRPRP
HAdV2 12S    PEIHPVVPLCPIKPVAVRVGGRRQAVECIEDLLNESGQPLDLSCKRPRP
             **************************************************
```

The sequence alignment difference between the HAdV2 12S and 13S E1A proteins is essentially the 46 residues corresponding to the CR3 domain, which is predicted to be folded. The CR3 domain is only present on the 13S E1A protein, between the residues 144 and 191.

Sequence alignment of HAdV 2/5 13S E1A proteins with HPV16 E7 protein

```
E1A      MRHIICHGGVITEEMAASLLDQLIEEVLADNLPPPSHFEPPTLHELYDLDVTAPEDPNEE
E7       ------------------------------------MHGDTPTLH----------------
                                              * : ****

E1A      AVSQIFPESVMLAVQEGIDLFTFPPAPGSPEPPHLSRQPEQPEQRALGPVSMPNLVPEVI
E7       -------------------------------------------------EYMLDLQPETT
                                                           * :* **.

E1A      DLTCHEAGFPPSDDEDEEGEE-----------FVLDYVGHPGHGCRSCH---------YH
E7       DLYCYEQLNDSSEEEDEIDGPAGQAEPDRAHYNIVTFCCKCDSTLRLCVQSTHVDIRTLE
         ** *:*     *::***                 :: :  :     * *           .

E1A      RRNTGDPDIMCSLCYMRTCGMFVYSPVSEPEPEPEPEPEPARPTRRPKLVPAILRRPTSP
E7       DLLMGTLGIVCPICSQKP-----------------------------------------
             *   *:* :*  :

E1A      VSRECNSSTDSCDSGPSNTPPEIHPVVPLCPIKPVAVRVGGRRQAVECIEDLLNEPGQPL
E7       ------------------------------------------------------------


E1A      DLSCKRPRP
E7       ---------
```

The alignment between the HAdV2/5 hybrid and the HPV 16 E7 protein showed an identity matrix of 29.5 %. The observed similarity is essentially between the CR1 and CR2 regions from both proteins.