Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/virology

Interplay between sequence, structure and linear motifs in the adenovirus E1A hub protein

Juliana Glavina[a], Ernesto A. Román[b], Rocío Espada[a], Gonzalo de Prat-Gay[c], Lucía B. Chemes[d,e,*], Ignacio E. Sánchez[a,**]

[a] Universidad de Buenos Aires. Consejo Nacional de Investigaciones Científicas y Técnicas. Instituto de Química Biológica de la Facultad de Ciencias Exactas y Naturales (IQUIBICEN). Facultad de Ciencias Exactas y Naturales. Laboratorio de Fisiología de Proteínas. Buenos Aires, Argentina
[b] Instituto de Química y Físico-Química Biológicas, Universidad de Buenos Aires, Junín 956, 1113AAD, Buenos Aires, Argentina
[c] Protein Structure-Function and Engineering Laboratory, Fundación Instituto Leloir and IIBBA-CONICET, Buenos Aires, Argentina
[d] Consejo Nacional de Investigaciones Científicas y Técnicas. Instituto de Investigaciones Biotecnológicas IIB-INTECH, Universidad Nacional de San Martín, San Martín, Buenos Aires, Argentina
[e] Departamento de Fisiología y Biología Molecular y Celular (DFBMC), Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina

ARTICLE INFO

Keywords:
Adenovirus
E1A
Intrinsic disorder
Linear motifs
Motif repertoire
Sequence conservation
Co-evolution
Random polymer

ABSTRACT

E1A is the main transforming protein in mastadenoviruses. This work uses bioinformatics to extrapolate experimental knowledge from Human adenovirus serotype 5 and 12 E1A proteins to all known serotypes. A conserved domain architecture with a high degree of intrinsic disorder acts as a scaffold for multiple linear motifs with variable occurrence mediating the interaction with over fifty host proteins. While linear motifs contribute strongly to sequence conservation within intrinsically disordered E1A regions, motif repertoires can deviate significantly from those found in prototypical serotypes. Close to one hundred predicted residue-residue contacts suggest the presence of stable structure in the CR3 domain and of specific conformational ensembles involving both short- and long-range intramolecular interactions. Our computational results suggest that E1A sequence conservation and co-evolution reflect the evolutionary pressure to maintain a mainly disordered, yet non-random conformation harboring a high number of binding motifs that mediate viral hijacking of the cell machinery.

1. Introduction

The *Adenoviridae* family groups ubiquitous small, non-enveloped DNA viruses with an icosahedral capsid Reddy et al., 2010). Over 250 characterized adenovirus serotypes from five genera infect a wide range of vertebrate hosts (Harrach et al., 2011), with all serotypes infecting humans belonging to the *Mastadenovirus* genus. Adenoviruses infect mucoepithelial cells and persist in the lymphatic system, occasionally infiltrating the cerebrospinal fluid and brain tissue within the central nervous system (Khanal et al., 2018). Adenovirus infection can lead to acute respiratory diseases, pneumonia, gastroenteritis, keratoconjunctivitis and acute hemorrhagic cystitis (Khanal et al., 2018). Because of their oncogenic properties, adenoviruses are often classified together with papillomaviruses and polyomaviruses as small dsDNA tumor viruses. All human adenoviruses can transform baby rat kidney cells *in vitro*, while Human adenovirus 12 (HAdV12) but not Human adenovirus 5 (HAdV5) is also able to cause tumors in immunocompetent rodents (Graham et al., 1977; Williams et al., 2004).

The adenovirus genome is a double stranded linear DNA molecule coding for 20–50 proteins (Davison et al., 2003). The genes transcribed early in the viral reproductive cycle code for proteins involved in virus-host interactions that enable viral genome replication and transcription (King et al., 2018). Among these, the E1A gene is unique to the genus *Mastadenovirus*, which infects a variety of mammals including humans (Davison et al., 2003). The E1A protein is essential for a productive viral infection, deregulating the host cell cycle and transcriptional machinery in favor of conditions suitable for viral replication (Pelka et al., 2008). In cooperation with the E1B protein, E1A transforms rodent cells *in vitro* and is a main oncogenicity determinant (Williams et al., 2004). The biological activity of the E1A protein involves the

[*] Corresponding author at:.Consejo Nacional de Investigaciones Científicas y Técnicas. Instituto de Investigaciones Biotecnológicas IIB-INTECH, Universidad Nacional de San Martín, San Martín, Buenos Aires, Argentina.
[**] Corresponding author.
E-mail addresses: lchemes@iib.unsam.edu.ar (L.B. Chemes), isanchez@qb.fcen.uba.ar (I.E. Sánchez).

https://doi.org/10.1016/j.virol.2018.08.012
Received 22 May 2018; Received in revised form 13 August 2018; Accepted 14 August 2018
Available online 25 September 2018

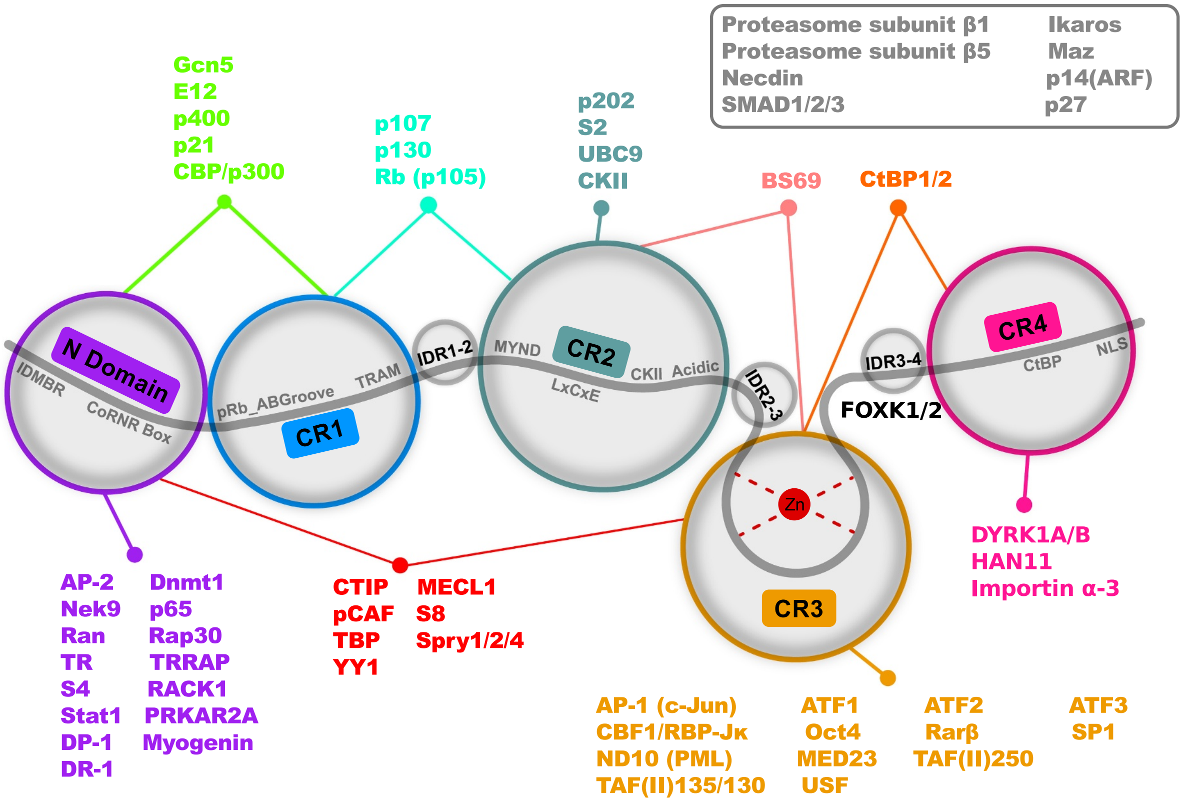

Fig. 1. Schematic representation of E1A oncoprotein domains and interactions. The approximate location of each conserved domain is indicated by colored circles (N domain, purple; CR1, blue; CR2, green; CR3, yellow; CR4, pink). The CR3-associated zinc atom is represented as a red sphere and the approximate location of the linear motifs defined in Table 1 is also shown, namely IDMBR, CoRNR Box, pRB_ABGroove, TRAM, MYND, LxCxE, CKII, Acidic Stretch, CtBP, NLS. The Inter Domain Regions IDR12, IDR23 and IDR34 described in the results sections are indicated by gray circles. The size of each circle is not related to the size of the domain or region. The targets for each domain of the E1A oncoprotein have been compiled from literature (File S1) and are shown grouped according to their target domain/ regions, with the unmapped targets are shown in gray.

formation of protein-protein interactions with more than 50 host target molecules (Fig. 1), making it a crucial molecular hub for infection. Well-studied E1A protein targets include the Retinoblastoma (pRb) tumor suppressor protein, p300 and CBP (Boyd et al., 2002; Whyte et al., 1988a; Ferreon et al., 2009), all of which are shared with small DNA tumor virus oncoproteins such as papillomavirus E7 and the polyomavirus Large T antigen. Among other effects, these interactions lead to relocalization and functional inactivation of the retinoblastoma protein family members (pRb, p130, and p107) and of the p300/CBP regulators of histone acetylation (Ferrari et al., 2008). Through this global perturbation in the cell transcriptional network (Ferrari et al., 2008), the E1A protein subverts the cell cycle and contributes to immune system evasion (Strath and Blair, 2006).

The sequence features of the 280-residue long E1A oncoprotein are related to its large number of molecular interactions (Pelka et al., 2008; King et al., 2018). The HAdV5 E1A gene encodes five protein isoforms (Radko et al., 2015). The two largest isoforms differ in an internal sequence of 46 residues (Perricaudet et al., 1979). The largest E1A consists of 5 domains, which are referred to in the literature as follows: N domain, CR1, CR2, CR3 and CR4 (Kimelman et al., 1985; Avvakumov et al., 2002). The CR3 domain is absent in the second largest isoform (Perricaudet et al., 1979). The CR1 and CR2 domains present similarity to sequence stretches within papillomavirus E7 and the polyomavirus Large T antigen. In addition to the canonical domains, a “spacer” was identified as an oncogenic determinant in HAdV12 E1A (Telling and Williams, 1994), and auxiliary regions 1 and 2 were characterized in HAdV5 E1A as co-regulators of viral early gene transcription (Bondesson et al., 1992). However, it is yet unknown whether these regions are present in other E1A proteins.

The structural features of E1A and its regions are still poorly understood. Nuclear magnetic resonance experiments indicated that the CR1, CR2 and CR4 domains of HAdV5 E1A are intrinsically disordered, while the N domain shows partial order and the CR3 domain folds into a poorly understood alpha-helical structure (Pelka et al., 2008; Hošek et al., 2016), but no X-ray diffraction structures have been reported for E1A or its domains. The CR3 domain contains a double CxxC motif involved in structural zinc binding (Culp et al., 1988) reminiscent of the globular domain of papillomavirus E7, although there is no discernible sequence similarity outside of the double CxxC motif. Structural studies using fragments from the HAdV5 and/or HAdV12 E1A proteins further confirmed that the CR1 (Ferreon et al., 2009), CR4 (Molloy et al., 2000) and the N-CR1 domains (Haberz et al., 2016) are disordered when isolated from the rest of the protein. E1A is intrinsically disordered yet able to undergo disorder-to-order transitions. For example, in the presence of trifluoroethanol the binding sites for the CtBP and TBP proteins adopt beta turn and alpha conformations, respectively (Molloy et al., 1998, 1999, 2000). Additionally, both the CR1 domain and the N domain-CR1 constructs undergo local disorder-to-order transitions upon binding of a target protein (Ferreon et al., 2009; Haberz et al., 2016). This folding-upon-binding phenomenon may regulate the simultaneous formation of multiple protein interactions (Ferreon et al., 2013).

Binding of E1A to many of its host cellular targets can be rationalized in terms of multiple linear motifs (Davey et al., 2011), which are short sequence elements of 5–15 residues often found within intrinsically disordered domains (Davey et al., 2012) that mediate many protein-protein interactions and are often key to host-pathogen interactions and molecular mimicry (Davey et al., 2011). Biochemical studies performed mainly on HAdV12 E1A or HAdV5 E1A showed that several E1A domains are densely packed with linear motifs mediating binding to host proteins (Fig. 1, Table 1). The N domain presents an

Table 1
Motif definitions. From left to right: E1A domain, motif and its target indicated between brackets, criteria for identification of instances, percentage of sequences with an instance of the motif, references for the first description of the motif and relevant updates, and E1A protein(s) where the motif was experimentally characterized. In regular expressions (Dinkel et al., 2014) wild card positions are indicated by a dot, which implies that any amino acid is allowed in that position or by [^xy] meaning that amino acids x and y are forbidden in that position. Fixed positions are indicated by [xy], meaning that amino acids x and y are allowed in that position or by one amino acid, meaning that only that amino acid is allowed in that position. The sign means "OR".

Domain	Motif (Target)	Regular Expression	% Sequences	References	E1A type
N Domain	**IDMBR** (S8, TATA Binding Protein, CREB Binding Protei(see Methodn)	I......T......LL..L....L.D |	6	Boyd et al. (2002)	HAdV5
		C...........LL..L.... L |		Rasti et al. (2005)	
		R...C....[IL][ST]......LL..L[IL]			HAdV5
	CoRNR Box (Thyroid Receptor)	[ILF][^P][^P][ILVFY][ILV] [^P][^P][^P][ILVYFHM]	89	Meng et al. (2005)	HAdV5
				Phelan et al. (2010)	–
CR1	**pRb_ABGroove** (Retinoblastoma)	[IVLA].[NQDE][IVLFMYA][IVLFMYA] [AHKTNQDES]	95	Chemes et al. (2012b)	–
				Ikeda and Nevins (1993)	HAdV5
				Dyson et al. (1992)	HAdV5
				Whyte et al. (1988b)	HAdV5
	TRAM-CBP (CREB Binding Protein)	[DE].[NQ][DE][^PG]AV[^PG] [NQDEST][ILMVF] F....[MIL][^PG] A[AV][^PG]..[IVLF]	9	Ferreon et al. (2009)	HAdV5
CR2	**MYND** (BS69)	P.L.P	52	Hateboer et al. (1995)	–
				Ansieau and Leutz (2002)	–
				Isobe et al. (2006)	–
				Dinkel et al. (2014)	–
	LxCxE (Retinoblastoma)	[IL].C.[DE]	97	Whyte et al. (1988a)	HAdV5
				Dyson et al. (1992)	HAdV5
				Ikeda and Nevins (1993)	HAdV5
				Corbeil and Branton (1994)	–
				Dinkel et al. (2014)	–
	Acidic region (Retinoblastoma)	Net charge ≤ -4	94	Chemes et al. (2012b)	–
	CKII (Casein Kinase II)	[ST]..[DE]	97	Allende and Allende (1995)	–
				Whalen et al. (1996)	–
CR3	**Cysteine Rich Positions**	% Cys ≥ 5.9	42	Chemes et al. (2012a)	–
				Chemes et al. (2014)	–
	Zinc binding motif	CxxC	100	Culp et al. (1988)	HAdV5
CR4	**NLS** (Importin α)	[^DE]K[RK][KRP][KR][^DE]	74	Lyons et al. (1987)	HAdV5
				Köhler et al. (2001)	–
				Madison et al. (2002)	HAdV5/2
				Dinkel et al. (2014)	–
	CtBP (C-Terminal Binding Protein)	P.DLS	75	Boyd et al. (1993)	HAdV5/2
				Schaeper et al. (1995)	HAdV5/2
				Molloy et al. (1998), Molloy et al. (2006), Molloy et al. (2007)	HAdV12
				Cohen et al. (2013)	HAdV5
				Dinkel et al. (2014)	–

intrinsically disordered multiple binding region (IDMBR) (Rasti et al., 2005; Boyd et al., 2002), and a CoRNR Box motif (Meng et al., 2005; Phelan et al., 2010). Next in sequence, the CR1 domain presents the pRb_ABGroove motif that confers binding to the Retinoblastoma protein AB domain (Dyson et al., 1992; Liu and Marmorstein, 2007) and the TRAM-CBP binding motif (Ferreon et al., 2009). The pRb_ABGroove motif is shared with the papillomavirus E7 protein. The CR2 domain contains a MYND domain binding motif (Hateboer et al., 1995; Isobe et al., 2006; Ansieau and Leutz, 2002), an LxCxE motif that binds to a different pocket in the Retinoblastoma protein AB domain (Whyte et al., 1988a; Dyson et al., 1992; Ikeda and Nevins, 1993; Corbeil and Branton, 1994), a Casein kinase II phosphorylation site (Whalen et al., 1996) and an acidic stretch that also cooperate with the LxCxE motif in binding to Retinoblastoma (Palopoli et al., 2018) and are shared with papillomavirus E7 and polyomavirus Large T antigen. E1A interacts with transcriptional regulators including binding sites for host transcription factors (Chatton et al., 1993; Liu and Green, 1994) and TBP-associated factors (Geisberg et al., 1995; Mazzarelli et al., 1997) within CR3 and a C-terminal Binding protein (CtBP) binding motif within CR4 (Boyd et al., 1993; Schaeper et al., 1995; Molloy et al., 1998, 2006, 2007; Cohen et al., 2013) next in sequence to a Nuclear Localization Signal (NLS) (Lyons et al., 1987; Köhler et al., 2001; Madison et al., 2002). In several cases such as the pRb protein and p300/CBP, multiple motifs act together in binding of the same cellular target.

To date, no systematic studies have assessed the evolutionary conservation of sequence, intrinsic disorder and linear motifs in the adenovirus E1A hub protein, making it not obvious whether the experiments performed on the HAdV5 and HAdV12 E1A proteins can be extrapolated to E1A proteins from other serotypes. Here, we take advantage of a wealth of available experimental and sequence information on the adenovirus E1A molecular hub to perform a comprehensive bioinformatics study (Chemes et al., 2012a) and propose E1A as a model system for understanding the conservation of different sequence positions within linear motifs (Davey et al., 2012) and the variability of linear motif repertoires within viral proteins (Chemes et al., 2015). Moreover, E1A emerges as a model for shedding light on poorly understood areas such as the relative conservation of intrinsically disordered regions compared to globular domains, the amount of long range order compatible with intrinsic disorder (Varadi et al., 2014) and the prediction of transient residue contacts from sequence information (Toth-Petroczy et al., 2016).

2. Results

2.1. Three novel inter domain regions in the adenovirus E1A protein

Molecular biology (Kimelman et al., 1985; Subramanian et al., 1988) and sequence analysis (Avvakumov et al., 2002) of prototypical mastadenovirus E1A proteins led to the definition of five conserved domains, designated N domain and CR1 to CR4. The boundaries for these domains were established using 34 human adenovirus E1A sequences (Avvakumov, 2004). We revisited this result using an up to

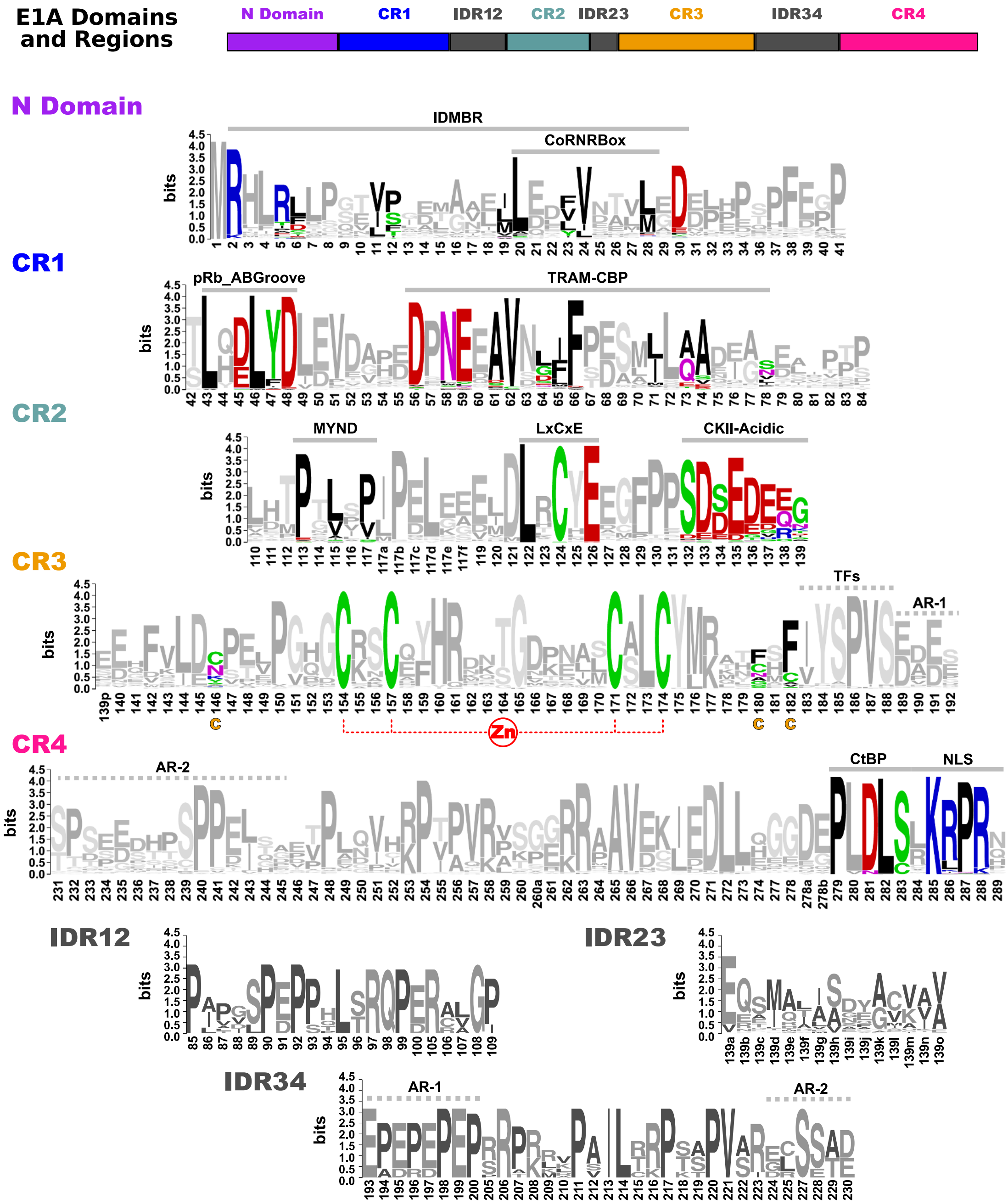

Fig. 2. Sequence conservation on the E1A oncoprotein. Based on the alignment of 116 mastadenovirus E1A sequences, regions of poor conservation were removed and positions that could be aligned with confidence are depicted as sequence logos (see Sequence database and alignment Methods section and source data from Files S2 and S3). Sequence logos represent the information content for each position in the alignment in units of bits, with the maximum information content value for protein sequences being 4.32 bits (see Methods section Sequence logos). For each domain, only the subset of sequences where that domain was conserved were considered. From top to bottom: N domain (113 sequences), CR1 (43 positions from 112 sequences), CR2 (114 sequences), CR3 (116 sequences), CR4 (107 sequences), IDR12 (12 sequences) and IDR23 (89 sequences), and IDR34 (12 sequences). Sequence numbering originates in the HAdV5 E1A protein. Insertions relative to the HAdV5 E1A sequence are indicated with a number followed by a letter. The positions of the known functional motifs are marked above the logos as gray lines. Fixed positions are shown in color and all other positions are shown in gray. The four CR3 cysteine residues involved in zinc binding are indicated with a dotted red line, three Cys-rich positions are indicated below the logo with a yellow C, and the binding site for transcription factors (TFs) are indicated at the top of the logo with a dotted gray line. Auxiliary regions 1 (AR-1) and 2 (AR-2) are indicated at the top of the logos with a dotted gray line.

date database of 116 mastadenovirus E1A sequences from 23 adenovirus species (see Methods for details). Previously reported domains were easily recognizable in the sequence alignment and were present in nearly all E1A sequences. We further refined the domain boundaries using sequence conservation, percentage of gaps and known functional motifs and defined a systematic sequence numbering that refers to the prototypical HAdV5 E1A protein (Fig. 2). The alignment does not present poorly aligned stretches (File S2 and File S3), and the known domains immediately follow each other in sequence in most cases. This is in contrast to previous results for the papillomavirus E7 protein, where the known domains and functional motifs were in some cases separated by poorly conserved and aligned sequence segments best described as linkers.

In addition, we identified three blocks of well-aligned positions that were not present in all sequences. We term these blocks inter domain regions (Fig. 2, bottom). Inter Domain Region 1–2 (IDR12) was located between CR1 and CR2 and may play a role in formation of a ternary complex with the retinoblastoma protein and CBP/p300 (Ferreon et al., 2009). IDR12 was present only in the 12 serotypes of *Human adenovirus C*, which includes the two prototypical serotypes HAdV2 and HAdV5. Inter Domain region 2–3 (IDR23) was identified between CR2 and CR3 and corresponds to the spacer reported as an oncogenic determinant in HAdV12 E1A from species *HA* (Williams et al., 2004). IDR23 was present in 89 E1A proteins from many primate and porcine adenovirus serotypes, namely in all sequences from species *HA*, *HB*, *HD*, *HE*, *HF*, *PA*, *SA*, *SB* and *SF*, and in one of the three E1A sequences from species *HG*. Last, Inter Domain Region 3–4 (IDR34) was present between CR3 and CR4. The first eight residues and the last 6 residues of this 33 residue region belong to the auxiliary regions 1 and 2 previously characterized in HAdV5 E1A from species *Human adenovirus C* (Bondesson et al., 1992). IDR34 was present in the same sequences as IDR12.

In conclusion, our comprehensive analysis of the 116 known E1A sequences confirmed the 5 known domains (N domain, CR1, CR2, CR3 and CR4) and updated their boundaries (Fig. 2). It also uncovered three inter domain regions present in 12 (IDR12 and IDR34) to 89 sequences (IDR23) (Fig. 2).

2.2. Interplay between sequence conservation and intrinsic disorder in E1A

We used sequence logos (Schneider et al., 1986) to summarize E1A domain organization and to measure sequence conservation at the domain/region and residue level (Fig. 2). In contrast with the common expectation for a protein with a high degree of intrinsic disorder, many sequence positions show medium to high conservation (information values of 2 bits or higher). We compared the average level of conservation of the eight E1A domains and interdomain regions by performing pairwise comparisons of average domain information values and assigning *p*-values to the observed differences using permutation tests (Figure S1, see Methods for details). E1A domains and inter domain regions could not be clearly separated in terms of conservation, although it is worth noting that the average conservation value of the N domain was significantly smaller than for the CR2 and CR4 domains (*p*-value <0.05). To sum up, the overall degree of sequence conservation in E1A is high and the CR1 to CR4 domains of E1A are as conserved as the newly identified inter domain regions.

Computational analysis of six prototypical E1A sequences suggested that CR1, CR2 and CR4 are intrinsically disordered, while the N domain is mainly disordered and CR3 is mainly globular (Pelka et al., 2008). Here, we test the generality of this result by studying the disorder propensity for our database of 116 sequences with the IUPRED algorithm (see Methods for details). The IUPRED score can take values between 0 and 1. A score above 0.5 indicates disorder and a score below 0.5 indicates order (Dosztányi et al., 2005; Daughdrill et al., 2011). These results are represented in Fig. 3, where the x axis indicates the sequence position and the y axis indicates the mean IUPRED score and standard deviation for confidently aligned positions in each domain/region.

We tested whether the average IUPRED score of each E1A domain/region was above or below 0.5, using bootstrapping to generate confidence intervals for the difference between the average IUPRED score and the threshold of 0.5 (see Methods for details). The average IUPRED scores of the CR1, CR2 and CR4 domains and the IDR12 and IDR34 regions were above 0.5, suggesting that these domains and regions were disordered (*p*-value <0.01). The average IUPRED score of IDR23 is indistinguishable from 0.5 in our test, with a disordered first half and an ordered second half. The average IUPRED scores of the N domain and the CR3 domain suggest that these domains are ordered (*p*-value <0.01). In the case of the N domain, the N terminus seems to be more ordered than the C terminus. This is in agreement with nuclear magnetic resonance experiments (Hošek et al., 2016) and with the proposed presence of an amphipatic alpha helix in the N domain of the HAdV5 E1A protein (Pelka et al., 2008). In the case of the CR3 domain, a globular central stretch appears to be flanked by more flexible tails. In summary, we confirmed that the CR1, CR2 and CR4 domains and the IDR12 and IDR34 regions are predicted to be intrinsically disordered, while the IDR23 region is predicted to be partially ordered and the N domain and the CR3 domain are predicted to be mainly globular. The low standard deviation values for most positions also suggest that both structure as well as intrinsic disorder are predicted to be conserved properties of E1A domains.

Intrinsically disordered domains are believed to be less conserved than globular ones (Daughdrill et al., 2011; Toth-Petroczy et al., 2008). However, this is not necessarily the case (Chemes et al., 2012a), specially for viral proteins. A position-by-position plot of the average IUPRED score versus bits of sequence conservation for the adenovirus E1A protein shows no correlation (Figure S2). Comparing the average conservation of each domain (Figure S1) with the average IUPRED score (Fig. 3) also fails to show a strong relationship between predicted disorder and conservation. Thus, intrinsic disorder alone does not appear to dictate a lower sequence conservation for E1A domains/regions.

2.3. Twelve E1A linear motifs with strongly different prevalences

Experimental studies have identified a dozen functional linear sequence motifs within E1A proteins from a few mastadenovirus serotypes, such as HAdV12 from species *Human adenovirus A* or HAdV2 and HAdV5 from species *Human adenovirus C*. Ten out of the twelve E1A linear motifs mediate protein-protein interactions with host proteins (see Introduction and Table 1). The remaining two are within CR3. First, four cysteine residues coordinate a zinc atom and are thought to drive formation of a folded structure in this domain. Second, previous studies on the papillomavirus E7 oncoprotein suggest that three Cys-rich positions in the E1A CR3 may regulate E1A activity in response to changes in the redox environment in the host cell (Chemes et al., 2012a, 2014; Camporeale et al., 2016).

We studied linear motif prevalence among the E1A protein sequences. For ten out of these twelve linear motifs we used the sequence definitions available in the literature, while for the IDMBR and the TRAM-CBP motif we combined the available experimental evidence into new definitions (Table 1, see Methods for details). We then determined the presence and absence of each motif in the different E1A sequences, restricting the search to the site of the protein where the motif was first described. Four protein-protein interaction motifs (pRb_ABGroove, LxCxE, the CKII phosphorylation site and the acidic stretch) were present in most sequences. Four additional protein-protein interaction motifs were present in 52–89% of sequences (CoRNR Box, MYND, CtBP and NLS). Also, two protein-protein interaction motifs (IDMBR and TRAM-CBP) were present in less than 10% of E1A sequences. The CR3 zinc binding motif was present in all E1A sequences, and 42% of the E1A sequences had at least one cysteine at one of the three Cys-rich positions. In conclusion, the known E1A linear motifs identified in prototypical E1A sequences cannot be easily extrapolated to other serotypes using our regular expressions. The regular expressions described here, which are novel for the IDMBR and the TRAM-CBP motifs, may guide future experimental studies in mastadenovirus E1A proteins.

2.4. E1A linear motifs contribute to protein sequence conservation

According to our computational definition of linear motifs, we can define three different classes of sequence positions in a linear motif (Fig. 4). Fixed positions show a small number of allowed amino acids and are shown in color in Fig. 2. On the other hand, wild-card positions of a motif allow any amino acid. Third, adjacent positions include the three positions before and after the linear motif. Finally, positions that do not belong to a known linear motif are termed “other” in our analysis. Wild card, adjacent and “other” positions are shown in gray in Fig. 2. We analyzed the degree of conservation in these four classes of

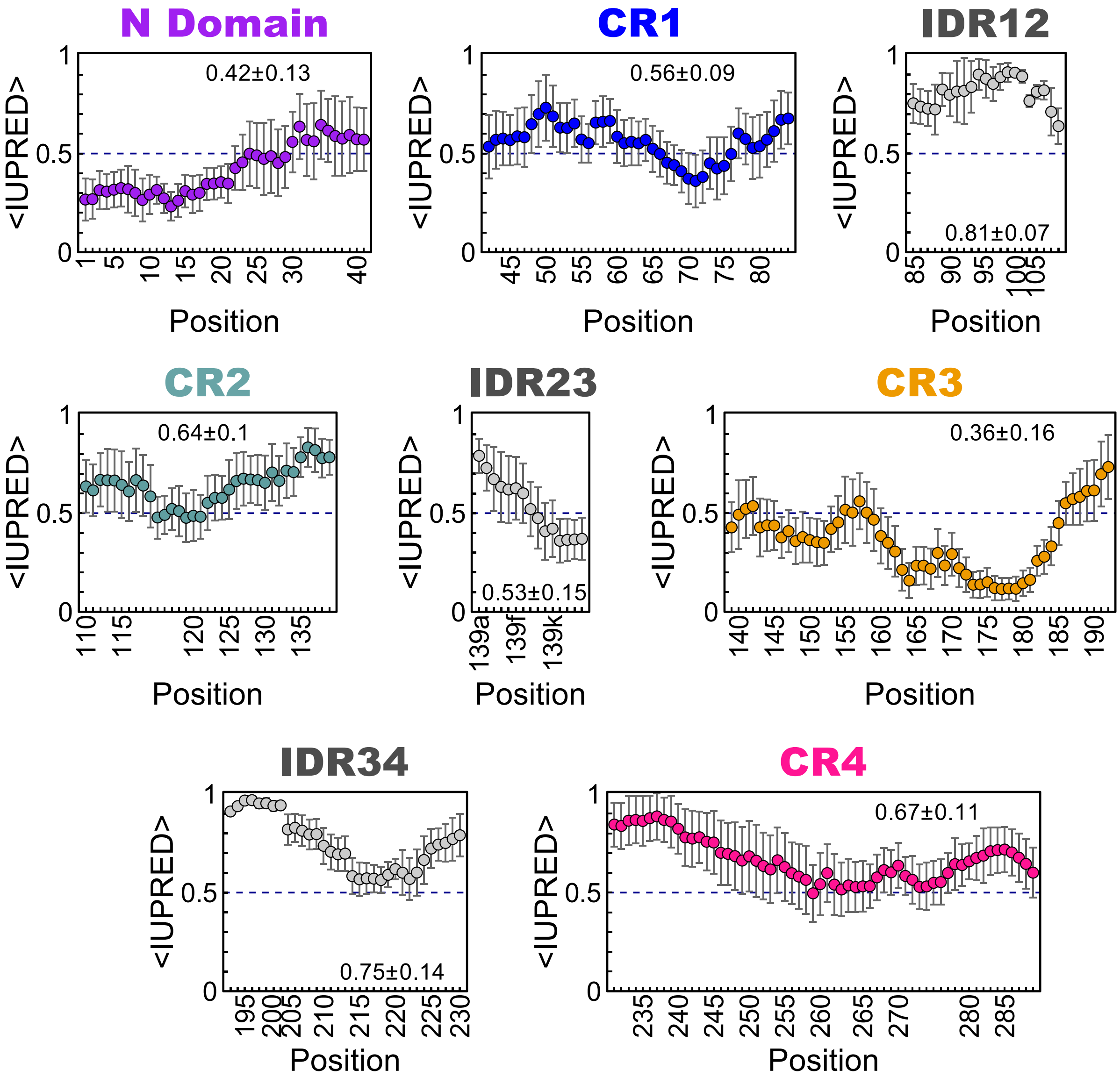

Fig. 3. Intrinsic disorder prediction for each domain and region of the E1A oncoprotein. The IUPRED algorithm was used to calculate a disorder score for all positions of all E1A sequences in our database. The score ranges from 0 for a fully ordered residue to 1 for a fully disordered residue. The results for the different E1A domains/ regions are displayed in different panels. Each panel shows the average IUPRED score and its standard deviation (y axis) for each position in the E1A alignment with less than 30% gaps (x axis). The source data are available as Supplementary File S5. Sequence numbering originates in the HAdV5 E1A protein. Insertions relative to the reference protein lead to non-correlative sequence numbers in some instances. The color code is the same used in Fig. 1.

sequence positions using information content as a measure of conservation (see Methods). The CKII phosphorylation site, the acidic region and Cys-rich positions were considered as “other” in this analysis because they do not act by themselves but rather modulate the activity of the LxCxE and CxxC motifs. The results are depicted as histograms in Fig. 4 (top row). To assess whether there was a significant difference between the average conservation values of the four classes, we calculated *p*-values using permutation tests and corrected them for multiple comparisons (see Methods). The average conservation value of the fixed sequence positions was significantly higher (*p*-value <0.05) than the average conservation value of the wild card, adjacent and “other” positions. No significant differences were observed between the average conservation value of the wild card sequence positions, adjacent positions and the rest of positions, or between the average conservation value of the wild card + adjacent sequence positions and the rest of positions. Thus, the twelve known E1A linear motifs lead to increased sequence conservation mainly at fixed sequence sites.

2.5. Diversity in the linear motif repertoire of natural E1A proteins

We analyzed the predicted motif repertoire of the 116 E1A sequences in our database (Fig. 5, left). E1A sequences are generally motif-rich, with 84% of sequences presenting 8 or more of the 12 known motifs. The twelve known E1A linear motifs may appear in 2^{12} combinations, yielding 4096 potential different motif repertoires for this protein. However, only 25 motif combinations are observed in natural E1A sequences and only three of them are present in more than five serotypes. The most abundant motif combination is present in 42 serotypes and includes the 9 most prevalent linear motifs. The six most prevalent linear motifs are present in the ten most abundant combinations. Thus, there is a correlation between motif prevalences and the E1A motif repertoires. We also investigated the phylogenetic distribution of E1A motif repertoires (Fig. 5, right). Linear motif combinations present in multiple serotypes are often present in multiple viral species, and several viral species with more than just one serotype in the database hold multiple linear motif combinations. This shows that

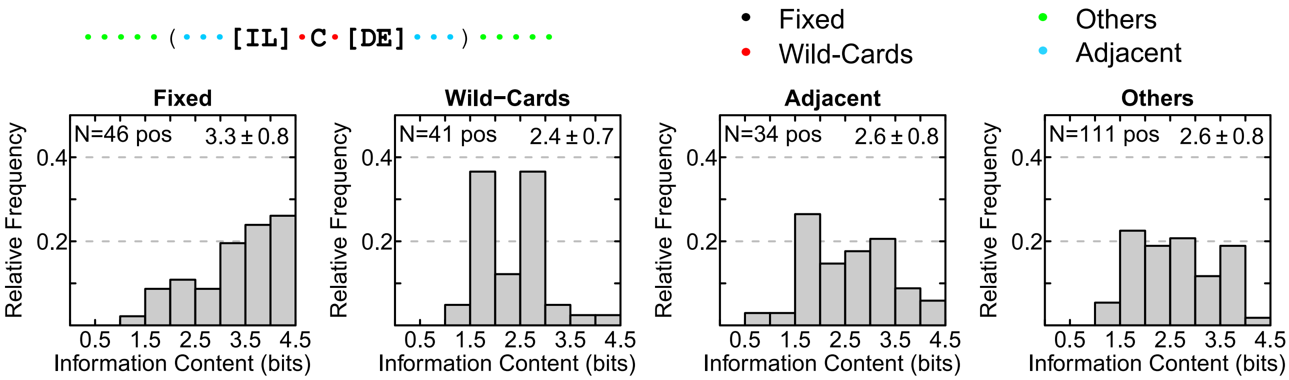

Fig. 4. Linear Motif sequence conservation and prevalence. Top: Diagram depicting the location of "Fixed", "Wild Card" and "Adjacent" positions relative to the regular expression of a representative linear motif. Bottom: The information content distribution for fixed, wild card and adjacent positions relative to each E1A motif and for all other positions within E1A is shown as histograms of relative frequency for different information content intervals. The number of positions analyzed and the mean value and standard deviation of the information content are displayed within each plot. Relative frequency has been computed as in the Methods section, using source data provided in Supplementary File S6. For more definitions, see the results section "E1A linear motifs contribute to protein sequence conservation".

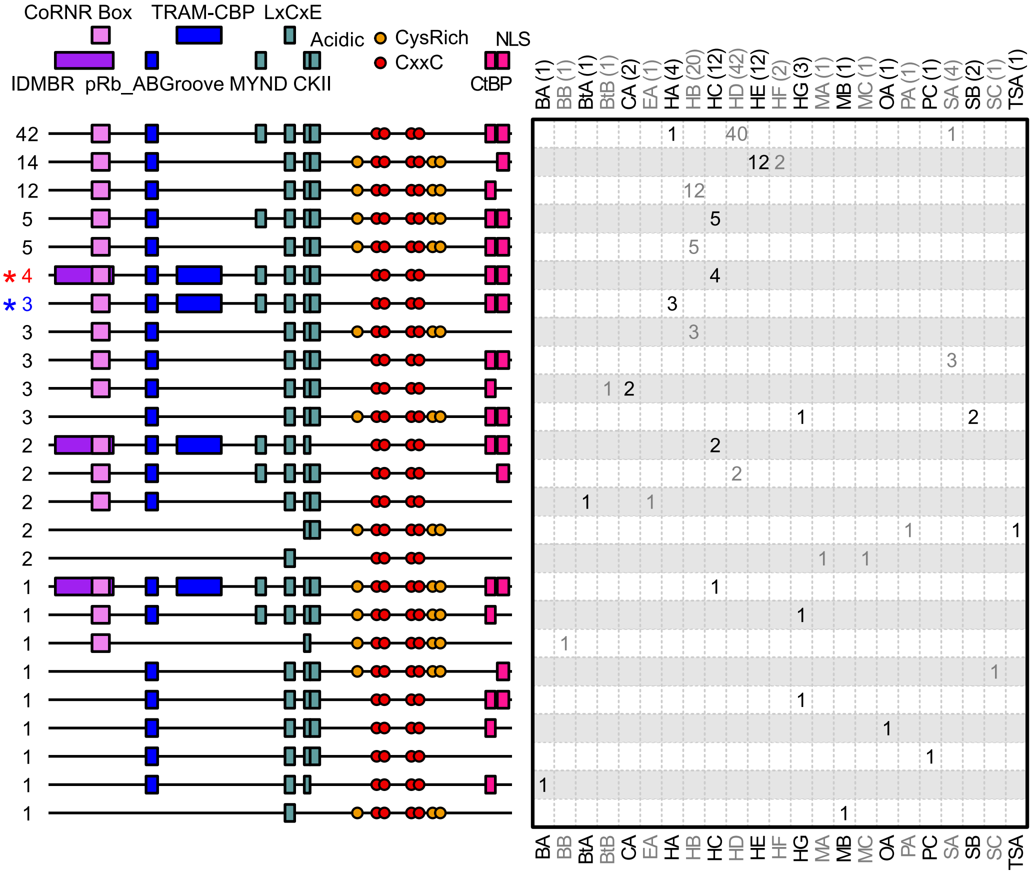

Fig. 5. Schematic representation of linear motif combinations repertoire. The 25 linear motif combinations found in the 116 E1A sequences are shown. All the motifs and their graphical representations are indicated at the top. The total number of sequences harboring the linear motif combination are indicated at the left. The linear motif combinations for HAdV5 and HAdV12 are indicated with red and blue stars respectively. The number of sequences and distribution within the adenovirus species are shown at the right. The abbreviations on the top of the right panel correspond to adenovirus species, as described in the Methods Section Sequence database and alignment, Table S1 and Supplementary File S7. The total number of sequences for each species are indicated between brackets next to the species name at the top. Linear motifs are represented as rectangles using the domain color code as in Fig. 1, except for the CoRNR Box (light pink rectangle) and the zinc binding motif (red dots). The size of each rectangle circle or line are not related to the size of the protein or linear motif.

adenovirus phylogeny does not completely determine the motif repertoire of the E1A protein. On the other hand, we observe a strong association between some motif repertoires and serotypes, such as the most abundant repertoire and species *HD*, suggesting specific linear motif evolutionary trends. Interestingly, the motif repertoires of the prototypical HAdV5 and HAdV12 E1A proteins are not on the top 5 most abundant repertoires, indicating that the most studied repertoires differ from those most abundant. In sum, the predicted E1A linear motif repertoire correlates with both motif prevalence and adenovirus phylogeny, but is not fully determined by these two factors.

2.6. Co-evolution of residues within disordered and globular domains of E1A

We investigated co-evolution of residue pairs in the E1A sequence alignment to gain insight on the sequence-structure relationships present in E1A. Co-evolution signals are a reliable indicator of conserved physical contact between residue pairs within a protein (Morcos et al., 2011). This is also the case for intrinsically disordered proteins (Toth-Petroczy et al., 2016). We used the Direct Information approach to infer residue contacts within E1A (see Methods). Our implementation of this method helps dissect direct and indirect co-evolution signals and takes into account aligned sequence redundancy (Espada et al., 2015).

Our computation identifies a total of 95 co-evolving residue pairs

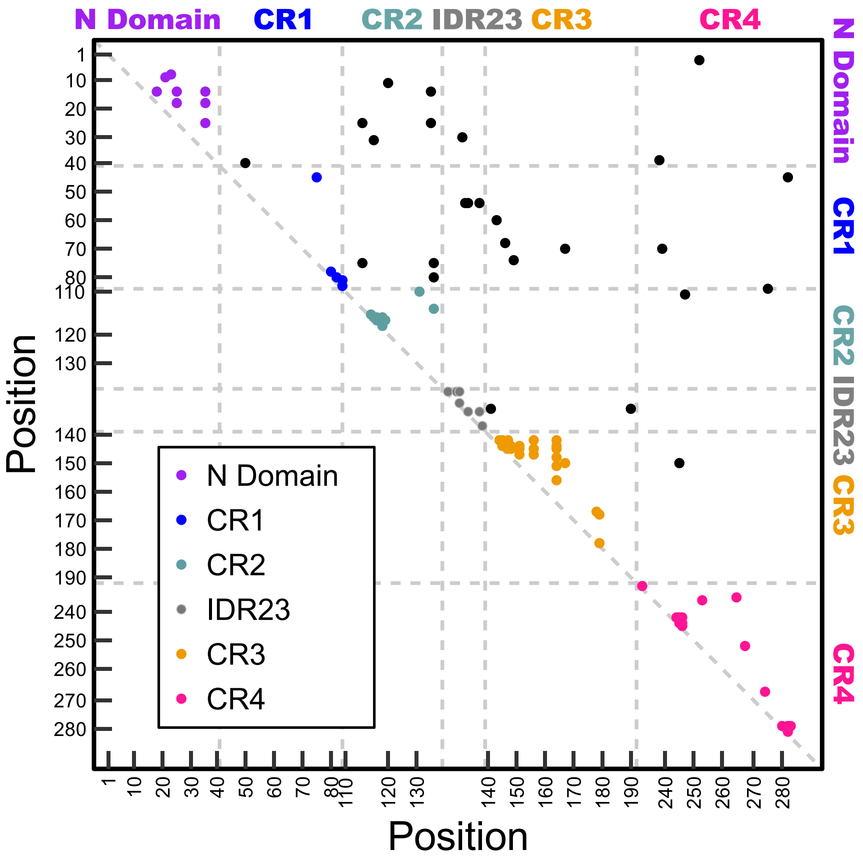

Fig. 6. Contact map prediction for the E1A protein using direct information, as defined and implemented in the Methods Section. The source data are available as Supplementary File S8. The E1A residue pairs with high values of direct information (Z-score ≥ 3) are represented as dots. Contacts within the same domain/region are indicated with the same color code used in Fig. 1. Contacts between different domain/regions are indicated in black. The different domains are separated by dashed lines. Sequence numbering originates in the HAdV5 E1A protein. Insertions relative to the reference protein lead to non-correlative sequence numbers in some instances. For better visualization, the contact predictions are displayed only above the diagonal. Regions IDR12 and IDR34 do not present residue pairs with high values of direct information.

within E1A. These pairs are represented in Fig. 6 as a predicted contact map, i.e., a two dimensional representation of the hits of direct information between amino acid pairs. 41 pairs were at 5 or less residues apart, 31 had a distance between 6 and 30 residues and the remaining 23 were 31 or more residues apart. The 69 co-evolving residue pairs belonging to the same domain are colored by domain in Fig. 6. Six E1A domains present intradomain co-evolving pairs, albeit at different frequencies. The CR3 and the IDR23 had the highest proportion of co-evolving pairs (~ 0.5 pairs/residue), with the CR2, the CR4 and the N domain presenting intermediate proportions between 0.2 and 0.3 pairs/residue and the CR1 presenting the lowest proportion at only ~ 0.12 pairs/residue. The remaining 26 co-evolving residue pairs involve different domains or regions, both ordered and disordered (Fig. 6, black circles). In sum, the co-evolution signals found here suggest a substantial amount of residue-residue contacts both within E1A domains and between them.

2.7. E1A protein is disordered, yet not a random chain

Disorder prediction suggests that E1A presents extensive regions with intrinsic disorder and smaller stretches of local order restricted to the CR3, the N domain and IDR23. On the other hand, the results from co-evolution point at tens of contacts between E1A residues that are far away in sequence. These two predictions seem to contradict each other, at least to some degree. We investigate whether the contacts predicted using the direct information method deviate from the expectations for a fully unstructured polypeptide, using an entropic chain model (Zhou, 2003, 2004). In this model, the only restrains to conformation are self-avoidance and the polymer persistence length (see Methods).

For each value of sequence distance within E1A, we calculated the apparent association contact between two aminoacids from the direct information data by comparing the number of predicted contacts to the number of all possible contacts ($K_{intramolecular}$ = number of residue pairs in contact/number of residue pairs not in contact). Since the number of predicted contacts is low, we averaged $K_{intramolecular}$ over a window of 14 residues (window size did not significantly alter the results). The results are shown as dots in Fig. 7. $K_{intramolecular}$ presents a minimum at around 50 residues of separation and grows at both smaller and larger sequence distances. Next, we tested the power of the entropic chain model to describe our results by fitting Eqs. (5) and (6) to the data. We obtained a value of $9.5{\cdot}10^{-5} \pm 0.8{\cdot}10^{-5}\,\mathrm{M}^{-1}$ for $K_{intermolecular}$ and of 6.1 ± 1.4 Angstroms for the $C_\alpha - C_\alpha$ contact distance. The latter value is in agreement with state-of-the-art analysis of residue contacts in globular proteins. However, while the theory describes the data well up to a sequence separation of about 25 residues, it consistently underestimates the observed $K_{intramolecular}$ values for sequence distances larger than 100 residues, leading us to conclude that an entropic chain model cannot account for the relatively high number of predicted long-range contacts between E1A residues. This suggests that E1A should not be described as a fully unstructured polypeptide.

2.8. A structural model for the E1A CR3 domain

Nuclear magnetic resonance (Ferreon et al., 2009; Hošek et al., 2016) and disorder prediction (Pelka et al., 2008 and this work) indicate that CR3 is the sole E1A domain that adopts a globular

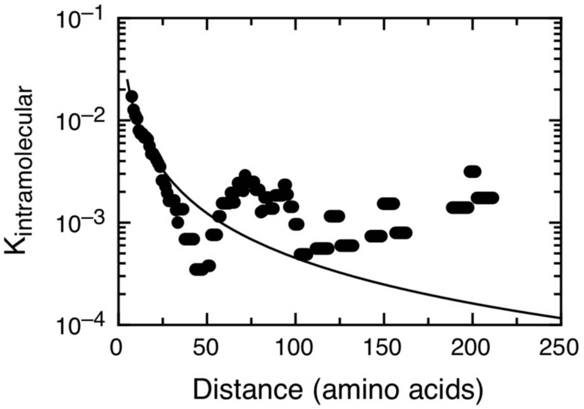

Fig. 7. Probability of contact formation in E1A as a function of sequence separation. Probability of contact formation in E1A as a function of sequence separation. The equilibrium constant for formation of contacts between amino acids $K_{intramolecular}$ = (number of residue pairs in contact / number of residue pairs not in contact) was estimated as a function of sequence separation L using the predicted contact map (dots). Since the number of predicted contacts is low, we averaged $K_{intramolecular}$ over a window of 14 residues (window size did not significantly alter the results) and these values were compared to those modeled considering E1A to behave as an entropic chain (Zhou, 2004). The best fit parameters for the equation describing the formation of intramolecular contacts in an entropic chain were a contact C_α–C_α distance of 6.1 ± 1.4 Angstroms, and an intermolecular constant $K_{intermolecular} = 9.5{\cdot}10^{-5} \pm 0.8{\cdot}10^{-5}$ mM^{-1}.

conformation, which is rich in alpha helices (Hošek et al., 2016). Building a homology model for the E1A CR3 domain was not possible because searching structure databases with the sequence of the CR3 domain did not yield any suitable template structure. Instead, we built an *ab initio* structural model for the CR3 domain using three kinds of information: residue contacts, secondary structure, and zinc binding. Our direct information analysis yielded a set of 14 high-confidence amino acid pairs within CR3 that are predicted to be in contact with each other (Fig. 8A, above the diagonal). Secondary structure prediction using the JPred algorithm suggests that residues 160–162 and 169–178 are in alpha helical conformation (Fig. 8A, top). The CR3 domain of E1A contains four cysteine residues that bind a zinc atom (Culp et al., 1988; Webster et al., 1991). The structure of the papillomavirus zinc binding domain suggested a close to planar Cys4 zinc binding site (Liu et al., 2006), with a distance of 10 Å between opposite cysteines and 7 Å between facing and close in sequence cysteines. This was used to include additional contact predictions in our model (Fig. 8B and triangles in Fig. 8A). An alternative arrangement of the four cysteines was also considered, with Cys157 facing Cys174 and Cys154 facing Cys171.

An initial unfolded structure of HAdV8 E1A CR3 was subjected to molecular dynamics simulations using a reduced C_α representation of the protein (Sułkowska et al., 2012) and the restrictions shown in Fig. 8A above the diagonal. The initial structure was subjected to heating/cooling cycles to unfold/collapse the protein. The zinc binding arrangement shown in Fig. 8B yielded a homogeneous set of annealed structures that satisfied most model constraints. Namely, 80% of the contact predictions are satisfied in the annealed structures (Fig. 8A, below the diagonal), which is typical for simulations of native state dynamics for domains of known structure. The distance between facing cysteines (black lines in Fig. 8B) and opposite cysteines (red lines in Fig. 8B) was consistent with the selected restraints. Also, 93% of the secondary structure prediction is satisfied at least 40% of the time (Fig. 8A, right). Fig. 8C shows the RMSD for the annealed ensemble, calculated using the average coordinates as a reference. We observe a well-defined globular structure in the N-terminal part of the domain (Fig. 8C, right), followed by a more flexible C-terminal segment (Fig. 8C, left). On the other hand, the alternative arrangement of the four cysteines failed to converge into a set of structures that satisfies most constraints simultaneously (data not shown). Thus, we chose the annealed structures resulting from the arrangement in Fig. 8B for further analysis.

We analyzed whether our model is compatible with our current knowledge of globular protein structures. We show in all cases the result for the same randomly chosen, representative member of the ensemble as in Fig. 8C. First, we analyzed the energetics of our structural model by measuring local configurational frustration. Typical globular domains consist of a minimally frustrated core with amino acid contacts that are energetically favorable in that structural context (Ferreiro et al., 2007). On the other hand, long loops and the protein surface may show patches of frustrated residue pairs that are in energetic conflict with the structure and often related to function (Ferreiro et al., 2007). Our CR3 model presents a minimally frustrated N-terminus that includes the zinc binding site (Fig. 8D, right) and a maximally frustrated C-terminus that is known to interact with host proteins (Fig. 8D, left). Second, we used the ConSurf algorithm (Ashkenazy et al., 2016) to evaluate sequence conservation at each position in the domain (Fig. 8E). The well-folded, minimally frustrated N-terminus shows a core of highly conserved positions, while the flexible, frustrated C-terminus is less conserved but for the residues involved in protein-protein interactions. Third, molecular dynamics simulations of experimentally determined structures are expected to display structural stability. In order to test our coarse grained structure prediction, we first reconstructed the amino acid side chains for a representative structure (See Methods). Structures were relaxed with a 2 ns of solvent-explicit molecular dynamics simulations. Fig. 8F shows that the values of RMSD per residue relative to the initial structure are mostly below 2 Å, which is typical for globular domains of this size.

Altogether, our structural model for the E1A CR3 domain shows energetic, conservation and structural stability patterns that are similar to those observed for experimental structures. We propose that the CR3 model may be useful in understanding the results of mutagenesis experiments on CR3 function and guide the search for an experimental structure, prevented so far by the structural heterogeneity of the CR3 domain in solution (Hošek et al., 2016).

3. Discussion

Molecular virology often focuses on performing detailed experiments on one or a few proteins from prevalent and/or clinically relevant serotypes. Although this is a sound choice, it leaves out a great deal of sequence, functional and host diversity that can provide insights into virus evolution and pathogenesis. Our approach is to find computational definitions of the features described for prototypical proteins that can be used for comprehensive sequence database search. The results presented here quantify conservation of each feature, test how far they can be extrapolated and suggest hypotheses for new experiments.

Five previously known E1A protein domains (N domain and CR1 to CR4) appear in most E1A sequences of our database (Fig. 2), a strict conservation supporting their functional significance. The “spacer” or IDR23 (Fig. 2) involved in E1A transforming activity (Larsen and Tibbetts, 1987; Telling and Williams, 1994; Doerfler and Böhm, 2004) is present in a subset of E1A proteins, suggesting that E1A regions that are not strictly conserved may provide clues towards adenovirus oncogenicity. In addition, two E1A regions are present in a small number of sequences (IDR12 and IDR34, Fig. 2). IDR34 is involved in transcriptional regulation and IDR12 in host protein binding, suggesting that some E1A molecular functions might have arisen throughout evolution as a result of domain gain/loss events.

The ten known E1A linear motifs do not extrapolate to uncharacterized sequences as well as the E1A domains (Table 1), with motif prevalence ranging from 6% to 100% (Table 1). Since E1A linear motifs mediate direct binding to host proteins, this predicted motif variability could lead to changes in E1A molecular interactions and

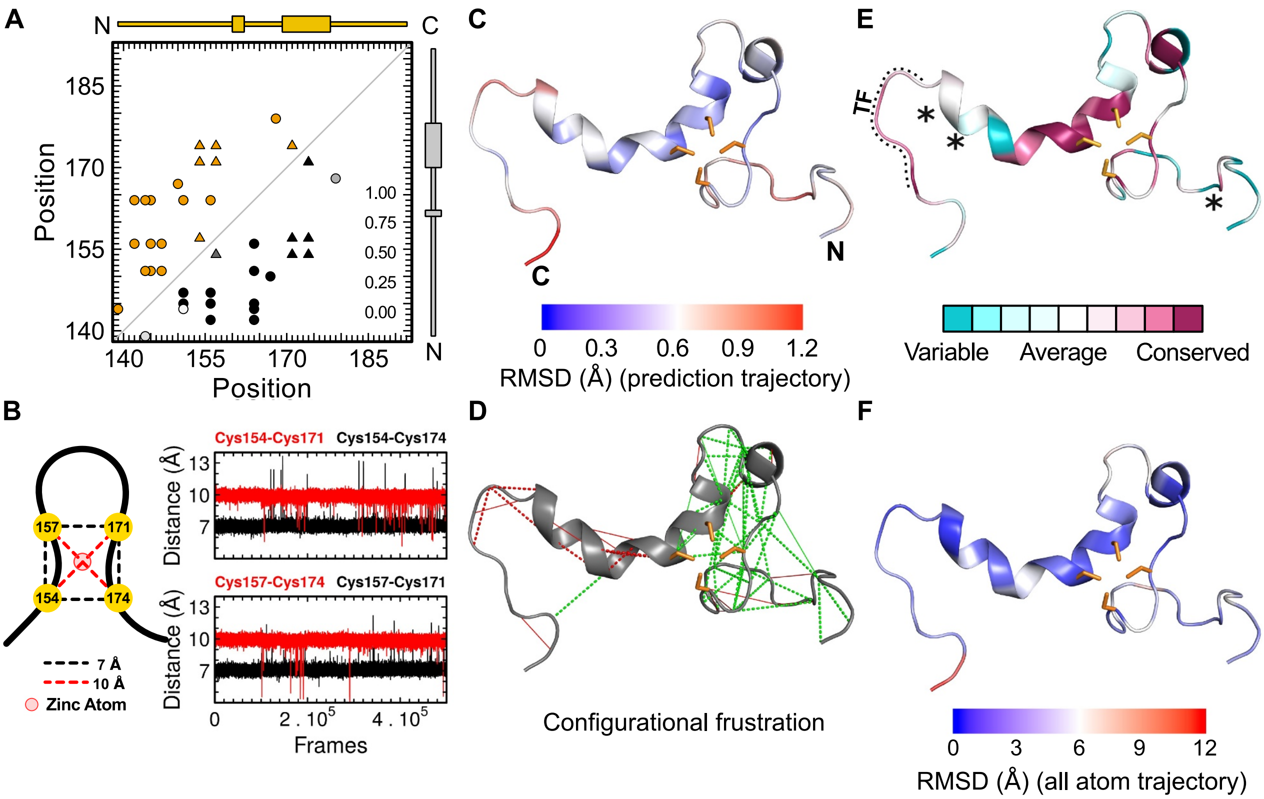

Fig. 8. A structural *de novo* model for the E1A CR3 domain. (A) Predicted Contact Map and Secondary Structure. Contacts used as restraints for structure prediction are represented as yellow dots (Upper half) or as dots colored by the probability of appearance on the annealed ensemble (Lower half). The six contacts of the cysteines involved in zinc coordination are represented as triangles. Secondary structure is represented at the top of the plot (JPred prediction) and at the right of the plot (annealed ensemble). Alpha helices are represented as rectangles. Sequence numbering originates in the HAdV5 E1A protein. (B) Configuration of the zinc binding site. The four cysteines involved in zinc coordination are represented as yellow circles. The interactions used as constraints in the simulation are represented as dotted lines. The zinc atom is represented as a light red circle. The two panels show the inter-cysteine distances along the simulation. (C) RMSD of predicted structures, using the average coordinates as reference and mapped over the protein backbone of a randomly chosen snapshot. Residues involved in zinc coordination are shown as yellow sticks. (D) Configurational Frustration. Highly frustrated contacts (red) and minimally frustrated contacts (green) are mapped over the protein backbone. Neutral contacts are not drawn. Residue-residue contacts are represented as solid lines and water mediated contacts are represented as dotted lines. (E) Evolutionary conservation of amino acid positions in CR3, from dark pink (most conserved) to light blue (least conserved). Cys-rich positions are indicated with black stars. The transcription factor binding site is indicated with a black dotted line. (F) Structural stability of the predicted CR3 structure in a molecular dynamics simulation. The RMSD values were calculated after a minimization and a short production simulation and mapped over the protein backbone of the initial structure.

function. Interestingly, motifs predicted to be of high prevalence such as the CoRNR Box and the Cys-rich positions continue to be discovered (Phelan et al., 2010 and this work) after twenty years of research on E1A (Dinkel et al., 2014). This, together with several highly conserved sequence positions with no known biological activity (Fig. 2), suggests that multiple E1A linear motifs await discovery, some of which might partially overlap with known motifs (Haberz et al., 2016). The most common predicted motif repertoire groups only 36% of E1A sequences (Fig. 5) and is distinct from that of the most widely studied serotypes, indicating that functional features discovered in prototypical sequences should not be assumed to appear in an archetypal E1A protein.

Most generic definitions of linear motifs do not account for differences between organisms (Haberz et al., 2016). We can now compare the specific E1A motif instances to generic regular expressions and to instances found in the papillomavirus E7 oncoprotein (Chemes et al., 2012a). Four motifs in E1A and E7 (pRb_ABGroove, LxCxE, acidic stretch and CKII sites) mediate binding to the host protein Retinoblastoma, and the last three constitute a functional module (Palopoli et al., 2018). In the LxCxE motif, a highly conserved aspartic acid (Palopoli et al., 2018) suggested to increase binding affinity to Retinoblastoma (Heck et al., 1992; Singh et al., 2005) preceeds the E1A motif (residue 121, Fig. 2). A fourth hydrophobic residue that can also increase affinity is conserved in the E1A motif three amino acids downstream of the motif, compared to two amino acids in E7 (Palopoli et al., 2018). The acidic stretch and phosphorylation of CKII sites can also enhance pRb binding affinity (Chemes et al., 2010, 2011). Both features are conserved in E1A, with a net charge of the acidic stretch ranging from − 4 to − 6 (Figure S3) and about half of E1A proteins presenting a single CKII site, while the other half presents two sites (Figure S3). Both numbers are similar to the ones found for E7 (unpublished data). The LxCxE flanking residues (Palopoli et al., 2018) and CKII and acidic modulatory elements are strikingly conserved in the E1A and E7 viral proteins despite their most likely independent origin, suggesting the convergent evolution of high affinity pRb binding mimicry in these viral proteins. Cys-rich positions in E7 relate to the redox behavior of the protein under different cellular conditions (Chemes et al., 2012a, 2014; Camporeale et al., 2016), which may also be the case for Cys-rich positions in E1A CR3. Compared to E7, less E1A sequences present non-canonical cysteines in this motif (Figure S3 and Chemes et al., 2012a) and a lower number of sequence positions are Cys-rich (Fig. 2 and Chemes et al., 2012a), suggesting a less prominent role for Cys-rich positions in E1A compared to E7. In conclusion, some linear motifs in E1A differ from generic definitions and from instances found in the E7 protein, supporting the view that the details of viral linear motif mimicry can be protein- or organism-specific.

We aimed at refining our coarse grained description of E1A structure (Fig. 1). As a first step, we looked for coevolution of residue pairs within this mainly disordered protein, as these signals correlate with residue contacts for both globular and intrinsically disordered proteins (Toth-Petroczy et al., 2016). We found predicted long-range contacts that deviate from the expectations for a fully disordered chain (Fig. 7) and extend to residue pairs separated up to 205 amino acids (Fig. 6). The results for E1A add to a growing body of evidence that intrinsically disordered proteins present non-random structural ensembles (Varadi et al., 2014). Future studies may help clarify the relationship between short-range (intradomain) and long-range (interdomain) order and protein-protein interactions mediated by multiple sequence stretches within E1A (Fig. 1). In particular, it will be interesting to correlate E1A structural properties to multiple linear motifs binding to the same protein, such as the LxCxE and pRB_ABGroove motifs targeting the retinoblastoma protein (Ferreon et al., 2009, 2013; Wright and Dyson, 2015).

We provide a testable structural model of the mainly globular CR3. The model is plausible in terms of residue contacts, zinc binding, energetics and sequence conservation (Fig. 8). We have used the TopMatch server (Sippl and Wiederstein, 2012) to look for similar structures in the protein data bank. As is often the case for viral proteins, we found no significant hits, suggesting that the E1A CR3 has no known structural homologs. The structural model also shows that the previously reported transcription factor binding site within CR3 is highly conserved yet loosely folded and energetically frustrated (Fig. 8). This suggests that the binding site for the ATF family, USF and Sp1 transcription factors and the TBP-associated factors $TAF_{II}250$ and $TAF_{II}135/130$ behaves as a partially independent module within a CR3 domain that is more complex than previously thought.

The sequence of the adenovirus E1A oncoprotein is highly conserved (Fig. 2 and Figure S1). Strikingly, conservation levels for interdomain regions present in only some E1A proteins are comparable to those of the CR1 to CR4 domains, which are present in all E1A sequences (Figure S1). Also, regions predicted to be intrinsically disordered (CR1, CR2, CR4, IDR12 and IDR34) are as conserved as the presumed partially ordered N domain and IDR23 or the globular CR3 (Figures S1 and Fig. 3), indicating that disordered regions can be as conserved as globular domains. The degree of predicted intrinsic disorder varies little across E1A sequences (Fig. 3), suggesting that disorder is an evolutionarily conserved structural property of E1A. We were able to relate E1A sequence conservation to the fixed sites of predicted linear motifs (Fig. 4). However, multiple E1A sequence positions outside of known functional motifs are highly conserved (Fig. 2), which may be due to several reasons. First, linear motifs awaiting discovery as discussed above, particularly in residues 231–278 from the CR4 domain and in the interdomain regions. One such motif may be a bipartite nuclear localization signal as suggested by Cohen et al. (2014), and another a binding site for the DREF transcription factor (Radko et al., 2014). Second, co-evolution with other E1A residues (Fig. 6) due to formation of protein complexes involving multiple sites within E1A (Fig. 1) (Dyson and Wright, 2016). Third, the avoidance of globular conformations (Borkosky et al., 2017), as with several E1A regions showing position-specific conservation of disorder-promoting aminoacids such as proline (Fig. 2). We propose that E1A sequence conservation reflects the evolutionary pressure to maintain a disordered, yet not random conformation as well as multiple binding motifs, which together allow the targeting of a high number of host proteins relative to the modest size of E1A (Fig. 1).

In conclusion, our comprehensive computational analysis of adenovirus E1A sequences provides a measure of how far individual experiments extrapolate to the E1A phylogeny, improving our understanding of the interplay between function, structure, disorder and sequence conservation that can guide future studies of this model viral oncoprotein.

4. Methods

4.1. Sequence database and alignment

We built a sequence database by retrieving 116 E1A sequences from the NCBI taxonomy database as of August 2012 (Supplementary File S7). These sequences are representative of the 23 species that belong to the *Mastadenovirus* genus within the adenovirus family and are the result of around 310 million years of viral evolution (Doerfler and Böhm, 2004). In order to keep a balanced representation across mastadenovirus serotypes, we retrieved a single reference E1A sequence for each serotype. The sequences and the corresponding serotype and species name abbreviations are listed in Table S1 and Supplementary File S7. All sequences were used to build an initial alignment using the MUSCLE software with default parameters (Edgar, 2004). The alignment was manually curated by taking into account sequence conservation at the known E1A functional sites. An additional degapped alignment was produced by removing 163 positions with more than 30% gaps from the curated alignment. Final alignments are available at File S3 and File S4.

The boundaries for the different E1A domains were previously defined in Avvakumov (2004) using 34 mastadenovirus E1A sequences. According to this, the N terminal domain encompasses the first 41 residues of HAdV5 E1A, CR1 spans residues 42–72, CR2 spans residues 115–137, CR3 spans residues 144–191 and CR4 spans residues 240–288. While no differences were observed for the N terminal domain in our alignment of 116 sequences, the boundaries of the four conserved regions were slightly different. The boundaries of CR1 shift from 72 to 84. We included the MYND linear motif within the boundaries of CR2. This expands CR2 to encompass 110–139. The boundaries of CR3 were easily identified at 140–192. Finally, the boundaries of CR4 were defined at 231–289 in HAdV5 E1A. Only 113 sequences sharing at least 20% identity with any sequence in the N-terminal region of the alignment were identified. 112 sequences were identified for the CR1 domain, 114 for the CR2 domain and 107 for the CR4 domain. The CR3 domain was conserved in all 116 sequences. Only the EAdV1 E1A and BAdV3 E1A lacked the CR4 domain.

The regions in the alignment between the CR1, CR2, CR3 and CR4 domains were used to define the inter domain regions IDR12, IDR23 and IDR34. Sequence stretches within these alignment regions that presented similar length and sequence were considered to belong to the relevant inter domain region. In the alignment region between the CR1 and CR2 domains, 57 sequences presented no amino acids, 42 sequences presented less than 10 residues, 14 sequences presented between 13 and 25 residues and the 3 remaining sequences showed a length over 29 residues. Out of the group of 14 sequences, 12 belonged to the *Human adenovirus C* species and shared at least 75% pairwise identity. These 12 sequences were used to define IDR12, with an average length of 22 ± 4 residues. In the alignment region between the CR2 and CR3 domains, 16 sequences presented no amino acids, 11 sequences presented less than 12 residues and the remaining 89 sequences presented between 14 and 25 residues. The group of 89 sequences belonged to the *Human adenovirus A*, *B*, *D*, *E*, *F* and *G*, *Porcine adenovirus A*, *Simian adenovirus A*, *B* and *F* species and shared at least 18% pairwise identity. These 89 sequences were used to define IDR23, with an average length of 16 ± 2 residues. In the alignment region between the CR3 and CR4 domains, 81 sequences presented no amino acids, 23 sequences presented less than 25 residues and the remaining 12 sequences presented between 32 and 38 residues. The group of 12 sequences all belonged to the species *Human adenovirus C* and shared at least 75% pairwise identity. These 12 sequences were used to define IDR34, with an average length of 35 ± 3 residues.

4.2. Sequence logos

Sequence logos (Schneider and Stephens, 1990) describing residue conservation for each E1A domain (N-domain, CR1, CR2, CR3 and CR4) and inter domain region (IDR12, IDR23 and IDR34) were generated using WebLogo (Crooks et al., 2004) and the corresponding degapped alignment of E1A. In a sequence logo, the x-axis indicates the position and the y-axis indicates the information content. The height of each stack indicates conservation of the position measured as information content in bits. The height of each letter inside the stack is proportional to the frequency of that letter in that position in the alignment. The letters are ordered according to the frequency inside each stack, with the most frequent at the top of each stack.

We calculated the information content $R(l)$ of each alignment position (Schneider and Stephens, 1990) using the degapped alignment of E1A. For each domain or region, we only included those E1A sequences where the domain/region was present. The information content was calculated as follows:

$$R(l) = log_2\ 20 - \left(-\sum_{b} f(b,\ l)\ log_2\ f(b,\ l)\right) - e(n) \tag{1}$$

where 20 is the alphabet size for proteins and $f(b,\ l)$ is the fraction of amino acid b at position l. *e(n)* is a small sample correction, where n is the number of sequences in the alignment. The maximum value of $R(l)$ is ≈ 4.32 ($\log_2 20$) bits, and the minimum value is zero.

4.3. Statistical analysis of amino acid conservation

The histograms of information content shown in Figure S1 and Fig. 4 were calculated as follows. The information content may attain values between zero and $\log_2 20$ bits (approximately 4.32 bits). We defined nine bins at intervals of 0.5 bits, going from 0 to 4.5 bits. We also defined sets of positions of the E1A full alignment (Supplementary File S2), such as domains and regions (Figure S1 and Supplementary File S4) or fixed/wildcard/adjacent/other positions relative to linear motifs (Fig. 4 and Supplementary File S6). We then counted, for each set of positions, how many confidently aligned positions of the E1A alignment (Supplementary File S3) had a value for the information content that falls within each bin. These frequencies were converted into relative frequencies by dividing by the total number of alignment positions considered for each E1A domain or set of positions and displayed in bar plots.

We tested for differences in the average amino acid conservation $R(l)$ of the different E1A domains and motif-related sets of positions. First, we used the Shapiro-Wilk test (Shapiro and Wilk, 1965) to assess the normality of the different $R(l)$ datasets. Since not all $R(l)$ datasets followed a normal distribution (p-value ≤ 0.05), we calculated p-values for differences in the average $R(l)$ values using permutation tests (Good, 2006). A permutation test is a non-parametric approach to establish the null distribution of a test statistic. Briefly, for two datasets A and B, with m and n observations respectively, we compute the statistic as:

$$\Theta = |\overline{X}_A - \overline{X}_B| \tag{2}$$

where $\overline{X}_A$ and $\overline{X}_B$ are the average conservation values for each data set. We combine the two sets of observations into one dataset with m +n observations. We randomly sampled without replacement sets of the same number of elements as A and B and compute the difference between the average values of the two sampled datasets. This step is performed 10,000 times. p-values are calculated as the fraction of times that the absolute value for the difference between the average values of the two sampled datasets is greater than or equal tothe absolute value of the observed difference out of a total number of permutations. Next, we applied the Benjamini-Hochberg correction (Benjamini and Hochberg, 1995) for multiple comparisons to control the false discovery rate. Briefly, the p-values obtained for each individual comparison were ordered in increasing order and an index i was assigned. The p-value was corrected as: $p_i^* = p\ \frac{m}{i}$, where p is the individual p-value and m is the total number of comparisons. Two $R(l)$ values datasets are considered to be different if the corrected p-value is smaller than the chosen 0.05 cutoff.

4.4. Derivation of regular expressions for the E1A linear motifs

Linear motifs in proteins can be formally described as regular expressions. In our analysis, we only included regular expressions of linear motifs with strong supporting evidence such as alanine scanning analysis, evidence of direct interaction *in vitro*, functionality analysis *in vivo* or crystallographic structure available. We defined the E1A linear motifs present in each domain of the protein as follows. The resulting definitions are listed in (Table 1).

N domain. The Intrinsically Disordered Multiple Binding Region (IDMBR) regular expression was defined according to the alanine scanning analysis and binding experiments performed on HAdV5 E1A by Rasti et al. (2005) and Boyd et al. (2002). We included in our definition those positions that when mutated decreased binding or activity by at least 50% and also conservative amino acid substitutions found in homologous proteins. The Corepressor nuclear receptor interaction motif (CoRNR Box motif) definition was modified from the regular expression proposed by Phelan et al. (2010) by adding conservative amino acid substitutions found in E1A sequences.

CR1. The regular expression for the pRb_ABGroove corresponds to the E2F mimic previously defined in Chemes et al. (2012b). The TRAM-CBP binding motif was defined according to the known structure (PDB:2KJE, Ferreon et al., 2009). We defined the E1A residues mediating CBP binding as those separated by a distance not greater than 5 Å from the CBP protein, having at least six contacts with CBP and having a RMSD lower than 1.3 Å in the 20 NMR models (Martin et al., 2010). In addition, we include all conservative amino acid substitutions found in E1A proteins.

CR2. The regular expressions for the MYND domain binding and LxCxE linear motifs were taken from Dinkel et al. (2014). The Casein kinase II phosphorylation site was defined by Allende and Allende (1995). In the case of the acidic stretch, we considered the sequence positions between the last position of the LxCxE motif and the end of the CR2 domain. It was considered as present if the net charge for the sequence positions under consideration was -4 or less (Chemes et al., 2012b).

CR3. The presence of Cys-rich positions in the CR3 domain of the E1A protein was defined as those confidently aligned positions with a cysteine abundance higher than 5.9% (Chemes et al., 2012a). The zinc binding motif consists of two copies of a CxxC subsequence 17 residues apart.

CR4. The regular expression for the CtBP binding linear motif is from Dinkel et al. (2014). The Nuclear Localization Signal was defined as the strong core version of the monopartite variant of the positively charged nuclear localization signal (TRG_NLS_MonoCore_2) (Dinkel et al., 2014).

4.5. Disorder prediction

We performed a position-by position prediction of the degree of intrinsic disorder for all sequences in our database using the algorithm IUPRED (Dosztányi et al., 2005), using the prediction type *long* disorder. IUPRED takes into account the ability of the sequence stretch surrounding an amino acid to form favorable residue-residue interactions.

We tested whether the average IUPRED score of the different E1A domains or regions is above or below 0.5. First, we used the Shapiro-Wilk test (Shapiro and Wilk, 1965) to assess normality of the different IUPRED score datasets. There was not enough evidence to reject the hypothesis that the datasets came from a population that has a normal distribution. However, a normal Q-Q plot comparing the dataset quantiles on the vertical axis to a normal population on the horizontal axis, suggested that the datasets were not normally distributed (see Figure S4). Hence, we used bootstrapping to generate 99% confidence intervals for the difference in the average IUPRED score and the threshold of 0.5. Briefly, we resampled the datasets with replacement 10,000 times and calculated the average for each bootstrap sample. We then sorted the 10,000 estimates as: $\bar{x}_1 \le \bar{x}_2 ... \bar{x}_{9999} \le \bar{x}_{10000}$. Consequently, the desired 99% bootstrap confidence interval is $(\bar{x}_{50}, \bar{x}_{9950})$.

4.6. Residue Co-evolution

The co-evolution of residue pairs can be inferred from the correlation of changes between two positions. One method to measure the correlation of amino acids changes between two positions of the alignment is using mutual information (Buslje et al., 2009), and a more recent and related implementation is direct information (Morcos et al., 2011). The direct information value is calculated as:

$$DI_{ij} = \sum_{A,B} P_{ij}^{dir}(A, B) \ln\left(\frac{P_{ij}^{dir}(A, B)}{f_i(A) f_j(B)}\right) \tag{3}$$

$P_{ij}^{dir}(A, B)$ is the probability that satisfies the constraint that the probability of the amino acid A at position i of the alignment $P_i(A)$, defined as the partial sum of the probability $P_{ij}^{dir}(A, B)$, matches the marginal frequency of the amino acid A at positions i and the probability of the amino acid B at position j of the alignment $P_j(B)$, defined as the partial sum of the probability $P_{ij}^{dir}(A, B)$ matches the marginal frequency of the amino acid B at positions j:

$$P_i(A) \equiv \sum_{B} P_{ij}^{dir}(A, B) = f_i(A) \tag{4a}$$

$$P_j(B) \equiv \sum_{A} P_{ij}^{dir}(A, B) = f_j(B) \tag{4b}$$

The direct information increases as the correlation between the positions gets higher. We chose to use direct information because, unlike mutual information, this method is capable of distinguishing between direct correlations of two amino acids and the correlation of two amino acids mediated by a third residue, providing insight into direct interactions. In this work, direct information was used to infer direct co-evolution among residue pairs in the degapped alignment of E1A according to Morcos et al. (2011). The phylogenetic bias in the collected sequences was corrected using heuristic sequence weights (Henikoff and Henikoff, 1994). To correct for spurious correlations generated by the finite number of sequences, we generated a site-independent alignment. Briefly, we individually scrambled each of the columns of the original alignment, keeping the marginal frequencies of the amino acids in each position but breaking all true correlations. We calculated direct information for the site-independent alignment and subtracted the results from the direct information score calculated on the original alignment (Espada et al., 2015). Finally, the direct information scores were translated into Z-scores and those pairs with Z ≥ 3 were selected for analysis.

4.7. Entropic chain model

As a model for a fully disordered protein, we used a previously described entropic chain model (Zhou, 2003) that considers the protein as a continuous rod that changes direction in a random manner (wormlike chain) with a constant radius of curvature. This allows us to model the probability for formation of a physical contact between two amino acids that are L residues apart in the protein by calculating an apparent equilibrium constant termed $K_{intramolecular}$. $K_{intramolecular}$ is the product of the equilibrium constant for association of two amino acids that are free in solution, $K_{intermolecular}$, and the entropic penalty for restricting the protein conformation to bring the two amino acids within contact distance. This entropic penalty term is called effective concentration or C_{eff}:

$$K_{intramolecular} = K_{intermolecular} {}^{*} C_{eff} \tag{5}$$

C_{eff} depends on three parameters: the persistence length of the linker l_p, the contour length l_c and the contact distance r. l_p is the length it takes for the changes in direction to become uncorrelated and in the case of proteins takes a value of 3 Angstroms. l_c is the length of the chain and amounts to 3.8 Angstroms per residue. r is the distance cutoff used to consider whether to amino acids in the protein are in physical contact. The empirical equation for C_{eff} has been validated in several model systems (Zhou, 2004; Borcherds et al., 2017) and can be written as follows:

$$\begin{aligned} C_{eff} = {} & \left(\frac{10^7}{6.022}\right)\left(\frac{3}{4\pi l_p l_c}\right)^{\frac{3}{2}} exp\left(\frac{-3r^2}{4 l_p l_c}\right) \\ & \left(1 - \frac{5 l_p}{4 l_c} + \frac{2r^2}{l_c^2} - \frac{33 r^4}{80 l_p l_c^3} - \frac{79 l_p^2}{160 l_c^2}\right. \\ & - \frac{329 r^2 l_p}{120 l_c^3} + \frac{6799 r^4}{1600 l_c^4} - \frac{3441 r^6}{2800 l_p l_c^5} \\ & \left. + \frac{1089 r^8}{12800 l_p^2 l_c^6}\right) \end{aligned} \tag{6}$$

This model has been shorn to account for the behavior of multiple unrelated unstructured protein linkers (Zhou, 2003, 2004).

4.8. CR3 structure prediction

We used a model where each residue is represented by a single bead centered on its alpha carbon position. Adjacent beads are strung together into a polymer chain by means of a potential encoding bond length and angle constraints through harmonic potentials. The secondary structure is encoded in the dihedral angle potential and the non-bonded (native contact) potential. Amino acid pairs with a high direct information score were considered to be present in the contact map of the native structure of the protein. Geometric predictions were made as in Sułkowska et al. (2012). Additional contact predictions were incorporated to define the zinc binding motif. Molecular dynamics simulations were performed using GROMACS 4.5.1 (Pronk et al., 2013) using simulated annealing protocol as in Sułkowska et al. (2012) and Clementi et al. (2000). A random initial structure of the protein was built using the CR3 sequence and Flexible Meccano (Ozenne et al., 2012) and subjected to cycles of heating/cooling. Structures collapsed in the freezing period were extracted for analysis. RMSD values were calculated using GROMACS 4.5 (Pronk et al., 2013) and VMD (Humphrey et al., 1996). Side chains of collapsed structures were rebuilt using Pulchra software (Rotkiewicz and Skolnick, 2008). Minimization and a short production simulation were done using TIP3P explicit water and AMBER99SB force field with a 2 fs step and Particle Mesh Ewald electrostatics. Protein representations were generated using Pymol (http://www.pymol.org).

Acknowledgements

We acknowledge funding from Agencia Nacional de Promoción Científica y Tecnológica (PICT 2012-2550 and PICT 2015-1213 to I.E.S. and PICT 2013-1895 to L.B.C.), Consejo Nacional de Investigaciones Científicas y Técnicas (Graduate fellowship to J.G. and R.E.; L.B.C., G.d.P.G., E.A.R. and I.E.S. are CONICET career investigators, L.B.C. is the recipient of PIP 2014-2016 11220130100558CO). Help from Alan Bush with statistical analyses is gratefully acknowledged. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.virol.2018.08.012.

References

Allende, J.E., Allende, C.C., 1995. Protein kinases. 4. Protein kinase CK2: an enzyme with multiple substrates and a puzzling regulation. FASEB J. 9 (5), 313–323. https://doi.org/10.1096/fasebj.9.5.7896000.

Ansieau, S., Leutz, A., 2002. The conserved Mynd domain of BS69 binds cellular and oncoviral proteins through a common PXLXP motif. J. Biol. Chem. 277 (7), 4906–4910. https://doi.org/10.1074/jbc.M110078200.

Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., et al., 2016. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res 44 (W1), W344–W350. https://doi.org/10.1093/nar/gkw408.

Avvakumov, N., Wheeler, R., D'Halluin, J.C., Mymryk, J.S., 2002. Comparative sequence analysis of the largest E1A proteins of human and simian adenoviruses. J. Virol. 76 (16), 7968–7975. https://doi.org/10.1128/JVI.76.16.7968-7975.2002.

Avvakumov, N., Kajon, a.E., Hoeben, R.C., Mymryk, J.S., 2004. Comprehensive sequence analysis of the E1A proteins of human and simian adenoviruses. Virology 329 (2), 477–492. https://doi.org/10.1016/j.virol.2004.08.007.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol. 57 (1), 289–300. https://doi.org/10.2307/2346101.

Bondesson, M., Svensson, C., Linder, S., Akusjärvi, G., 1992. The carboxy-terminal exon of the adenovirus E1A protein is required for E4F-dependent transcription activation. EMBO J. 11 (9), 3347–3354.

Borcherds, W., Becker, A., Chen, L., Chen, J., Chemes, L.B., Daughdrill, G.W., 2017. Optimal affinity enhancement by a conserved flexible linker controls p53 mimicry in MdmX. Biophys. J. 112 (10), 2038–2042. https://doi.org/10.1016/j.bpj.2017.04.017.

Borkosky, S.S., Camporeale, G., Chemes, L.B., Risso, M., Noval, M.G., Sánchez, I.E., et al., 2017. Hidden structural codes in protein intrinsic disorder. Biochemistry 56 (41), 5560–5569. https://doi.org/10.1021/acs.biochem.7b00721.

Boyd, J.M., Subramanian, T., Schaeper, U., La Regina, M., Bayley, S.T., Chinnadurai, G., 1993. A region in the C-terminus of adenovirus 2/5 E1a protein is required for association with a cellular phosphoprotein and important for the negative modulation of T24-ras mediated transformation, tumorigenesis and metastasis. EMBO J. 12 (2), 469–478.

Boyd, J.M., Loewenstein, P.M., Tang, Q.q., Yu, L., Green, M., 2002. Adenovirus E1A N-terminal amino acid sequence requirements for repression of transcription in vitro and in vivo correlate with those required for E1A interference with TBP-TATA complex formation. J. Virol. 76 (3), 1461–1474. https://doi.org/10.1128/JVI.76.3.1461-1474.2002.

Buslje, C.M., Santos, J., Delfino, J.M., Nielsen, M., 2009. Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. Bioinformatics 25 (9), 1125–1131. https://doi.org/10.1093/bioinformatics/btp135.

Camporeale, G., Lorenzo, J.R., Thomas, M.G., Salvatierra, E., Borkosky, S.S., Risso, M.G., et al., 2017. Degenerate cysteine patterns mediate two redox sensing mechanisms in the papillomavirus E7 oncoprotein. Redox Biol. 11, 38–50. https://doi.org/10.1016/j.redox.2016.10.020. (October 2016).

Chatton, B., Bocco, J.L., Gaire, M., Hauss, C., Reimund, B., Goetz, J., et al., 1993. Transcriptional activation by the adenovirus larger E1a product is mediated by members of the cellular transcription factor ATF family which can directly associate with E1a. Mol. Cell Biol. 13 (1), 561–570. https://doi.org/10.1128/MCB.13.1.561.

Chemes, L.B., Sánchez, I.E., Smal, C., 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 (4), 973–988. https://doi.org/10.1111/j.1742-4658.2009.07540.x.

Chemes, L.B., Sánchez, I.E., de Prat-Gay, G., 2011. Kinetic recognition of the retinoblastoma tumor suppressor by a specific protein target. J. Mol. Biol. 412 (2), 267–284. https://doi.org/10.1016/j.jmb.2011.07.015.

Chemes, L.B., Glavina, J., Alonso, L.G., Marino-Buslje, C., de Prat-Gay, G., Sánchez, I.E., 2012a. Sequence evolution of the intrinsically disordered and globular domains of a model viral oncoprotein. PLoS One 7 (10), e47661. https://doi.org/10.1371/journal.pone.0047661.

Chemes, L.B., Glavina, J., Faivovich, J., de Prat-Gay, G., Sánchez, I.E., 2012b. Evolution of linear motifs within the papillomavirus E7 oncoprotein. J. Mol. Biol. 422 (3), 336–346. https://doi.org/10.1016/j.jmb.2012.05.036.

Chemes, L.B., Camporeale, G., Sánchez, I.E., de Prat-Gay, G., Alonso, L.G., 2014. Cysteine-rich positions outside the structural zinc motif of human papillomavirus E7 provide conformational modulation and suggest functional redox roles. Biochemistry 53 (10), 1680–1696. https://doi.org/10.1021/bi401562e.

Chemes, L.B., de Prat-Gay, G., Sánchez, I.E., 2015. Convergent evolution and mimicry of protein linear motifs in host-pathogen interactions. Curr. Opin. Struct. Biol. 32, 91–101. https://doi.org/10.1016/j.sbi.2015.03.004.

Clementi, C., Nymeyer, H., Onuchic, J.N., 2000. Topological and energetic factors: what determines the structural details of the transition state ensemble and en-route intermediates for protein folding? An investigation for small globular proteins. J. Mol. Biol. 298 (5), 937–953. https://doi.org/10.1006/jmbi.2000.3693. (arXiv:0003460).

Cohen, M.J., Yousef, A.F., Massimi, P., Fonseca, G.J., Todorovic, B., Pelka, P., et al., 2013. Dissection of the C-terminal region of E1A redefines the roles of CtBP and other cellular targets in oncogenic transformation. J. Virol. 87 (18), 10348–10355. https://doi.org/10.1128/JVI.00786-13.

Cohen, M.J., King, C.R., Dikeakos, J.D., Mymryk, J.S., 2014. Functional analysis of the C-terminal region of human adenovirus E1A reveals a misidentified nuclear localization signal. Virology 468–470C, 238–243. https://doi.org/10.1016/j.virol.2014.08.014.

Corbeil, H.B., Branton, P.E., 1994. Functional importance of complex formation between the retinoblastoma tumor suppressor family and adenovirus E1A proteins as determined by mutational analysis of E1A conserved region 2. J. Virol. 68 (10), 6697–6709.

Crooks, G.E., Hon, G., Chandonia, J.m., Brenner, S.E., 2004. WebLogo: a sequence logo generator. Genome Res 14 (6), 1188–1190. https://doi.org/10.1101/gr.849004.

Culp, J.S., Webster, L.C., Friedman, D.J., Smith, C.L., Huang, W.j., Wu, F.Y., et al., 1988. The 289-amino acid E1A protein of adenovirus binds zinc in a region that is important for trans-activation. Proc. Natl. Acad. Sci. USA 85 (17), 6450–6454. https://doi.org/10.1073/pnas.85.17.6450.

Daughdrill, G.W., Borcherds, W.M., Wu, H., 2011. Disorder predictors also predict backbone dynamics for a family of disordered proteins. PLoS One 6 (12), 0–6. https://doi.org/10.1371/journal.pone.0029207.

Davey, N.E., Travé, G., Gibson, T.J., 2011. How viruses hijack cell regulation. Trends Biochem. Sci. 36 (3), 159–169. https://doi.org/10.1016/j.tibs.2010.10.002.

Davey, N.E., Van Roey, K., Weatheritt, R.J., Toedt, G., Uyar, B., Altenberg, B., et al., 2012. Attributes of short linear motifs. Mol. Biosyst. 8 (1), 268–281. https://doi.org/10.1039/c1mb05231d.

Davison, A.J., Benko, M., Harrach, B., 2003. Genetic content and evolution of adenoviruses. J. Gen. Virol. 84 (11), 2895–2908. https://doi.org/10.1099/vir.0.19497-0.

Dinkel, H., Van Roey, K., Michael, S., Davey, N.E., Weatheritt, R.J., Born, D., et al., 2014. The eukaryotic linear motif resource ELM: 10 years and counting. Nucleic Acids Res. 42 (Database issue), D259–66. https://doi.org/10.1093/nar/gkt1047.

Doerfler, W., Böhm, P., 2004. Adenoviruses: Model and Vectors in Virus-Host Interactions; vol. 273 of Current Topics in Microbiology and Immunology. 1 ed.; Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-05715-1. http://dx.doi.org/10.1007/978-3-662-05599-1.

Dosztányi, Z., Csizmók, V., Tompa, P., Simon, I., 2005. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21 (16), 3433–3434. https://doi.org/10.1093/bioinformatics/bti541.

Dyson, H.J., Wright, P.E., 2016. Role of intrinsic protein disorder in the function and interactions of the transcriptional coactivators CREB-binding Protein (CBP) and p300. J. Biol. Chem. 291 (13), 6714–6722. https://doi.org/10.1074/jbc.R115.692020. (arXiv:NIHMS150003).

Dyson, N.J., Guida, P., McCall, C., Harlow, E., 1992. Adenovirus E1A makes two distinct contacts with the retinoblastoma protein. J. Virol. 66 (7), 4606–4611.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32 (5), 1792–1797. https://doi.org/10.1093/nar/gkh340.

Espada, R., Parra, R.G., Mora, T., Walczak, A.M., Ferreiro, D.U., 2015. Capturing coevolutionary signals inrepeat proteins. BMC Bioinforma. 16 (1), 207. https://doi.org/10.1186/s12859-015-0648-3.(arXiv:1407.6903).

Ferrari, R., Pellegrini, M., Horwitz, G.a., Xie, W., Berk, A.J., Kurdistani, S.K., 2008. Epigenetic reprogramming by adenovirus e1a. Science 321 (5892), 1086–1088. https://doi.org/10.1126/science.1155546.

Ferreiro, D.U., Hegler, J.A., Komives, E.A., Wolynes, P.G., 2007. Localizing frustration in native proteins and protein assemblies. Proc. Natl. Acad. Sci. USA 104 (50), 19819–19824. https://doi.org/10.1073/pnas.0709915104.

Ferreon, J.C., Martinez-Yamout, M.A., Dyson, H.J., Wright, P.E., 2009. Structural basis for subversion of cellular control mechanisms by the adenoviral E1A oncoprotein. Proc. Natl. Acad. Sci. USA 106 (32), 13260–13265. https://doi.org/10.1073/pnas.0906770106.

Ferreon, A.C.M., Ferreon, J.C., Wright, P.E., Deniz, A.a., 2013. Modulation of allostery by protein intrinsic disorder. Nature 498 (7454), 390–394. https://doi.org/10.1038/nature12294.

Geisberg, J.V., Chen, J.L., Ricciardi, R.P., 1995. Subregions of the adenovirus E1A transactivation domain target multiple components of the TFIID complex. Mol. Cell. Biol. 15 (11), 6283–6290. https://doi.org/10.1128/MCB.15.11.6283.

Good, P.I., 2006. Permutation, Parametric, and Bootstrap Tests of Hypotheses. Springer Series in Statistics, 3 ed. Springer-Verlag, New York. https://doi.org/10.1111/j.1469-1809.2010.00572.x. (ISBN 9780387271583).

Graham, F.L., Smiley, J., Russell, W.C., Nairn, R., 1977. Characteristics of a human cell line transformed by DNA from human adenovirus type 5. J. Gen. Virol. 36 (1), 59–74. https://doi.org/10.1099/0022-1317-36-1-59.

Haberz, P., Arai, M., Martinez-Yamout, M.A., Dyson, H.J., Wright, P.E., 2016. Mapping the interactions of adenoviral E1A proteins with the p160 nuclear receptor coactivator binding domain of CBP. Protein Sci. 25 (12), 2256–2267. https://doi.org/10.1002/pro.3059.

Harrach, B., Benkö, M., Both, G.W., Brown, M., Davison, A.J., Echavarría, M., et al. 2011. In: King, A.M., Adams, M.J., Carstens, E.B., Lefkowitz, E.J., editors. Virus Taxononmy. Ninth Report. International Committe Taxon. Viruses; chap. Family Adenoviridae; 9 ed. San Diego: Elsevier. ISBN 9780123846846; pp. 125–141. http://dx.doi.org/10.1016/B978-0-12-384684-6.00009-4.

Hateboer, G., Gennissen, A., Ramos, Y.F.M., Kerkhoven, R.M., Sonntag-Buck, V., Stunnenberg, H.G., et al., 1995. BS69, a novel adenovirus E1A-associated protein that inhibits E1A transactivation. EMBO J. 14 (13), 3159–3169.

Heck, D.V., Yee, C.L., Howley, P.M., Münger, K., 1992. Efficiency of binding the retinoblastoma protein correlates with the transforming capacity of the E7 oncoproteins of the human papillomaviruses. Proc. Natl. Acad. Sci. USA 89 (10), 4442–4446. https://doi.org/10.1073/pnas.89.10.4442.

Henikoff, S., Henikoff, J.G., 1994. Position-based sequence weights. J. Mol. Biol. 243 (4), 574–578. https://doi.org/10.1016/0022-2836(94)90032-9.

Hošek, T., Calçada, E.O., Nogueira, M.O., Salvi, M., Pagani, T.D., Felli, I.C., et al., 2016. Structural and dynamic characterization of the molecular hub early Region 1A (E1A) from human adenovirus. Chem. - A Eur. J. 22 (37), 13010–13013. https://doi.org/10.1002/chem.201602510.

Humphrey, W., Dalke, A., Schulten, K., 1996. VMD: visual molecular dynamics. J. Mol. Graph 14 (1). https://doi.org/10.1016/0263-7855(96)00018-5. (33-8, 27-8).

Ikeda, M.A., Nevins, J.R., 1993. Identification of distinct roles for separate E1A domains in disruption of E2F complexes. Mol. Cell. Biol. 13 (11), 7029–7035. https://doi.org/10.1128/MCB.13.11.7029.

Isobe, T., Uchida, C., Hattori, T., Kitagawa, K., Oda, T., Kitagawa, M., 2006. Ubiquitin-dependent degradation of adenovirus E1A protein is inhibited by BS69. Biochem. Biophys. Res. Commun. 339 (1), 367–374. https://doi.org/10.1016/j.bbrc.2005.11.

028.

Köhler, M., Görlich, D., Hartmann, E., Franke, J., 2001. Adenoviral E1A protein nuclear import is preferentially mediated by importin alpha3. Vitr. Virol. 289 (2), 186–191. https://doi.org/10.1006/viro.2001.1151.

Khanal, S., Ghimire, P., Dhamoon, A., 2018. The repertoire of adenovirus in human disease: the innocuous to the deadly. Biomedicines 6 (1), 30. https://doi.org/10.3390/biomedicines6010030.

Kimelman, D., Miller, J.S., Porter, D., Roberts, B.E., 1985. E1a regions of the human adenoviruses and of the highly oncogenic simian adenovirus 7 are closely related. J. Virol. 53 (2), 399–409.

King, C.R., Zhang, A., Tessier, T.M., Gameiro, S.F., Mymryk, J.S., 2018. Hacking the cell: network intrusion and exploitation by adenovirus E1A. MBio 9 (3), 1–18. https://doi.org/10.1128/mBio.00390-18.

Larsen, P.L., Tibbetts, C., 1987. Adenovirus E1A gene autorepression: revertants of an E1A promoter mutation encode altered E1A proteins. Proc. Natl. Acad. Sci. USA 84 (December), 8185–8189. https://doi.org/10.1073/pnas.84.23.8185.

Liu, F., Green, M.R., 1994. Promoter targeting by adenovirus E1a through interaction with different cellular DNA-binding domains. Nature 368 (6471), 520–525. https://doi.org/10.1038/368520a0.

Liu, X., Marmorstein, R., 2007. Structure of the retinoblastoma protein bound to adenovirus E1A reveals the molecular basis for viral oncoprotein inactivation of a tumor suppressor. Genes Dev. 21 (21), 2711–2716. https://doi.org/10.1101/gad.1590607.

Liu, X., Clements, A., Zhao, K., Marmorstein, R., 2006. Structure of the human Papillomavirus E7 oncoprotein and its mechanism for inactivation of the retinoblastoma tumor suppressor. J. Biol. Chem. 281 (1), 578–586. https://doi.org/10.1074/jbc.M508455200.

Lyons, R.H., Ferguson, B.Q., Rosenberg, M., 1987. Pentapeptide nuclear localization signal in adenovirus E1a. Mol. Cell. Biol. 7 (7), 2451–2456. https://doi.org/10.1128/MCB.7.7.2451.

Madison, D.L., Yaciuk, P., Kwok, R.P.S., Lundblad, J.R., 2002. Acetylation of the adenovirus-transforming protein E1A determines nuclear localization by disrupting association with importin-alpha. J. Biol. Chem. 277 (41), 38755–38763. https://doi.org/10.1074/jbc.M207512200.

Martin, A.J.M., Walsh, I., Tosatto, S.C.E., 2010. MOBI: a web server to define and visualize structural mobility in NMR protein ensembles. Bioinformatics 26 (22), 2916–2917. https://doi.org/10.1093/bioinformatics/btq537.

Mazzarelli, J.M., Mengus, G., Davidson, I., Ricciardi, R.P., 1997. The transactivation domain of adenovirus E1A interacts with the C terminus of human TAF(II)135. J. Virol. 71 (10), 7978–7983.

Meng, X., Webb, P., Yang, Y.f., Shuen, M., Yousef, A.F., Baxter, J.D., et al., 2005. E1A and a nuclear receptor corepressor splice variant (N-CoRI) are thyroid hormone receptor coactivators that bind in the corepressor mode. Proc. Natl. Acad. Sci. USA 102 (18), 6267–6272. https://doi.org/10.1073/pnas.0501491102.

Molloy, D.P., Milner, A.E., Yakub, I.K., Chinnadurai, G., Gallimore, P.H., Grand, R.J.A., 1998. Structural determinants present in the C-terminal binding protein binding site of adenovirus early Region 1A proteins. J. Biol. Chem. 273 (33), 20867–20876. https://doi.org/10.1074/jbc.273.33.20867.

Molloy, D.P., Smith, K.J., Milner, A.E., Gallimore, P.H., Grand, R.J.A., 1999. The structure of the site on adenovirus early Region 1A responsible for binding to TATA-binding protein determined by NMR spectroscopy. J. Biol. Chem. 274 (6), 3503–3512. https://doi.org/10.1074/jbc.274.6.3503.

Molloy, D.P., Barral, P.M., Bremner, K.H., Gallimore, P.H., Grand, R.J.A., 2000. Structural determinants in adenovirus 12 E1A involved in the interaction with C-terminal binding protein 1. Virology 277 (1), 156–166. https://doi.org/10.1006/viro.2000.0580.

Molloy, D.P., Mapp, K.L., Webster, R., Gallimore, P.H., Grand, R.J.a., 2006. Acetylation at a lysine residue adjacent to the CtBP binding motif within adenovirus 12 E1A causes structural disruption and limited reduction of CtBP binding. Virology 355 (2), 115–126. https://doi.org/10.1016/j.virol.2006.05.004.

Molloy, D.P., Barral, P.M., Gallimore, P.H., Grand, R.J.a., 2007. The effect of CtBP1 binding on the structure of the C-terminal region of adenovirus 12 early region 1A. Virology 363 (2), 342–356. https://doi.org/10.1016/j.virol.2007.01.039.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., et al., 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc. Natl. Acad. Sci. USA 108 (49), E1293–301. https://doi.org/10.1073/pnas.1111471108.

Ozenne, V., Bauer, F., Salmon, L., Huang, J.R., Jensen, M.R., Segard, S., et al., 2012. Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. Bioinformatics 28 (11), 1463–1470. https://doi.org/10.1093/bioinformatics/bts172.

Palopoli, N., González Foutel, N.S., Gibson, T.J., Chemes, L.B., 2018. Short linear motif core and flanking regions modulate retinoblastoma protein binding affinity and specificity. Protein Eng. Des. Sel. 31 (3), 69–77. https://doi.org/10.1093/protein/gzx068.

Pelka, P., Ablack, J.N.G., Fonseca, G.J., Yousef, A.F., Mymryk, J.S., 2008. Intrinsic structural disorder in adenovirus E1A: a viral molecular hub linking multiple diverse processes. J. Virol. 82 (15), 7252–7263. https://doi.org/10.1128/JVI.00104-08.

Perricaudet, M., Akusjärvi, G., Virtanen, A., Pettersson, U., 1979. Structure of two spliced mRNAs from the transforming region of human subgroup C adenoviruses. http://dx.doi.org/10.1038/281694a0.

Phelan, C.A., Gampe, R.T., Lambert, M.H., Parks, D.J., Montana, V., Bynum, J., et al., 2010. Structure of Rev-erb alpha bound to N-CoR reveals a unique mechanism of nuclear receptor-co-repressor interaction. Nat. Struct. Mol. Biol. 17 (7), 808–814. https://doi.org/10.1038/nsmb.1860.

Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., et al., 2013. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics 29 (7), 845–854. https://doi.org/10.1093/bioinformatics/btt055.

Radko, S., Koleva, M., James, K.M.D., Jung, R., Mymryk, J.S., Pelka, P., 2014. Adenovirus E1A targets the DREF nuclear factor to regulate virus gene expression, DNA replication, and growth. J. Virol. 88 (22), 13469–13481. https://doi.org/10.1128/JVI.02538-14.

Radko, S., Jung, R., Olanubi, O., Pelka, P., 2015. Effects of adenovirus type 5 E1A isoforms on viral replication in arrested human cells. PLoS One 10 (10), 1–18. https://doi.org/10.1371/journal.pone.0140124.

Rasti, M., Grand, R.J.a., Mymryk, J.S., Gallimore, P.H., Turnell, A.S., 2005. Recruitment of CBP/p300, TATA-binding protein, and S8 to distinct regions at the N terminus of adenovirus E1A. J. Virol. 79 (9), 5594–5605. https://doi.org/10.1128/JVI.79.9.5594-5605.2005.

Reddy, V.S., Natchiar, S.K., Stewart, P.L., Nemerow, G.R., 2010. Crystal structure of human adenovirus at 3.5 Å resolution. Science (80-) 329 (5995), 1071–1075. https://doi.org/10.1126/science.1187292.

Rotkiewicz, P., Skolnick, J., 2008. Fast procedure for reconstruction of full-atom protein models from reduced representations. J. Comput. Chem. 29 (9), 1460–1465. https://doi.org/10.1002/jcc.20906.

Schaeper, U., Boyd, J.M., Verma, S., Uhlmann, E., Subramanian, T., Chinnadurai, G., 1995. Molecular cloning and characterization of a cellular phosphoprotein that interacts with a conserved C-terminal domain of adenovirus E1A involved in negative modulation of oncogenic transformation. Proc. Natl. Acad. Sci. USA 92 (23), 10467–10471. https://doi.org/10.1073/pnas.92.23.10467.

Schneider, T.D., Stephens, R.M., 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 18 (20), 6097–6100. https://doi.org/10.1093/nar/18.20.6097.

Schneider, T.D., Stormo, G.D., Gold, L., Ehrenfeucht, A., 1986. Information content of binding sites on nucleotide sequences. J. Mol. Biol. 188 (3), 415–431. https://doi.org/10.1016/0022-2836(86)90165-8.

Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for normality (Complete Samples). Biometrika 52 (3), 591–611. https://doi.org/10.1093/biomet/52.3-4.591.

Singh, M., Krajewski, M., Mikolajka, A., Holak, T.A., 2005. Molecular determinants for the complex formation between the retinoblastoma protein and LXCXE sequences. J. Biol. Chem. 280 (45), 37868–37876. https://doi.org/10.1074/jbc.M504877200.

Sippl, M.J., Wiederstein, M., 2012. Detection of spatial correlations in protein structures and molecular complexes. Structure 20 (4), 718–728. https://doi.org/10.1016/j.str.2012.01.024.

Strath, J., Blair, G.E., 2006. Adenovirus subversion of immune surveillance, apoptotic and growth regulatory pathways: a Model for tumorigenesis. Acta Microbiol. Immunol. Hung. 53 (2), 145–169. https://doi.org/10.1556/AMicr.53.2006.2.3.

Subramanian, T., Kuppuswamy, M., Nasr, R.J., Chinnadurai, G., 1988. An N-terminal region of adenovirus E1a essential for cell transformation and induction of an epithelial cell growth factor. Oncogene 2 (2), 105–112.

Sułkowska, J.I., Morcos, F., Weigt, M., Hwa, T., Onuchic, J.N., 2012. Genomics-aided structure prediction. Proc. Natl. Acad. Sci. USA 109 (26), 10340–10345. https://doi.org/10.1073/pnas.1207864109.

Telling, G.C., Williams, J., 1994. Constructing chimeric type 12/type 5 adenovirus E1A genes and using them to identify an oncogenic determinant of adenovirus type 12. J. Virol. 68 (2), 877–887.

Toth-Petroczy, A., Meszaros, B., Simon, I., Dunker, A.K., Uversky, V.N., Fuxreiter, M., 2008. Assessing conservation of disordered regions in proteins. Open Proteom. J. 1 (1), 46–53. https://doi.org/10.2174/1875039700801010046.

Toth-Petroczy, A., Palmedo, P., Ingraham, J., Hopf, T.A., Berger, B., Sander, C., et al., 2016. Structured states of disordered proteins from genomic sequences. Cell 167 (1), 158–170. https://doi.org/10.1016/j.cell.2016.09.010. (e12).

Varadi, M., Kosol, S., Lebrun, P., Valentini, E., Blackledge, M., Dunker, A.K., et al., 2014. PE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. Nucleic Acids Res. 42 (D1), 326–335. https://doi.org/10.1093/nar/gkt960.

Webster, L.C., Zhang, K., Chance, B., Ayene, I., Culp, J.S., Huang, W.j., et al., 1991. Conversion of the E1A Cys4 zinc finger to a nonfunctional His2,Cys2 zinc finger by a single point mutation. Proc. Natl. Acad. Sci. USA 88 (22), 9989–9993. https://doi.org/10.1073/pnas.88.22.9989.

Whalen, S.G., Marcellus, R.C., Barbeau, D., Branton, P.E., 1996. Importance of the Ser-132 phosphorylation site in cell transformation and apoptosis induced by the adenovirus type 5 E1A protein. J. Virol. 70 (8), 5373–5383.

Whyte, P., Buchkovich, K.J., Horowitz, J.M., Friend, S.H., Raybuck, M., Weinberg, R.A., et al., 1988a. Association between an oncogene and an anti-oncogene: the adenovirus E1A proteins bind to the retinoblastoma gene product. Nature 334 (6178), 124–129. https://doi.org/10.1038/334124a0.

Whyte, P., Ruley, H.E., Harlow, E., 1988b. Two regions of the adenovirus early region 1A proteins are required for transformation. J. Virol. 62 (1), 257–265.

Williams, J., Zhang, Y., Williams, M.A., Hou, S., Kushner, D., Ricciardi, R.P., 2004. E1A-based determinants of oncogenicity in human adenovirus groups A and C. Curr. Top. Microbiol Immunol. 273, 245–288. https://doi.org/10.1007/978-3-662-05599-1_8.

Wright, P.E., Dyson, H.J., 2015. Intrinsically Disordered Proteins in Cellular Signaling and Regulation. Nat. Rev. Mol. Cell Biol. 16 (1), 18–29. https://doi.org/10.1038/nrm3920.

Zhou, H.X., 2003. Quantitative account of the enhanced affinity of two linked scFvs specific for different epitopes on the same antigen. J. Mol. Biol. 329 (03), 1–8. https://doi.org/10.1016/S0022-2836(03)00372-3.

Zhou, H.X., 2004. Polymer Models of Protein Stability, Folding, and Interactions. Biochemistry 43 (8), 2141–2154. https://doi.org/10.1021/bi036269n.