Biochemistry

Subscriber access provided by MINCyT

Article

Hidden Structural Codes in Protein Intrinsic Disorder

Silvia S. Borkosky, Gabriela Camporeale, Lucia B. Chemes, Marikena Risso, M. Gabriela Noval, Ignacio Enrique Sánchez, Leonardo Gabriel Alonso, and Gonzalo de Prat-Gay

Biochemistry, **Just Accepted Manuscript** • DOI: 10.1021/acs.biochem.7b00721 • Publication Date (Web): 27 Sep 2017

Downloaded from http://pubs.acs.org on September 28, 2017

Just Accepted

“Just Accepted” manuscripts have been peer-reviewed and accepted for publication. They are posted online prior to technical editing, formatting for publication and author proofing. The American Chemical Society provides “Just Accepted” as a free service to the research community to expedite the dissemination of scientific material as soon as possible after acceptance. “Just Accepted” manuscripts appear in full in PDF format accompanied by an HTML abstract. “Just Accepted” manuscripts have been fully peer reviewed, but should not be considered the official version of record. They are accessible to all readers and citable by the Digital Object Identifier (DOI®). “Just Accepted” is an optional service offered to authors. Therefore, the “Just Accepted” Web site may not include all articles that will be published in the journal. After a manuscript is technically edited and formatted, it will be removed from the “Just Accepted” Web site and published as an ASAP article. Note that technical editing may introduce minor changes to the manuscript text and/or graphics which could affect content, and all legal disclaimers and ethical guidelines that apply to the journal pertain. ACS cannot be held responsible for errors or consequences arising from the use of information contained in these “Just Accepted” manuscripts.

ACS Publications

Biochemistry is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Published by American Chemical Society. Copyright © American Chemical Society. However, no copyright claim is made to original U.S. Government works, or works produced by employees of any Commonwealth realm Crown government in the course of their duties.

Hidden Structural Codes in Protein Intrinsic Disorder

Silvia S. Borkosky[†], Gabriela Camporeale[†], Lucía B. Chemes[†], Marikena Risso[†], María Gabriela Noval[§], Ignacio E. Sánchez[‡], Leonardo G. Alonso[†], and Gonzalo de Prat Gay[†]*

[†]*Protein Structure-Function and Engineering Laboratory, Fundación Instituto Leloir and Instituto de Investigaciones Bioquímicas de Buenos Aires (IIBBA) CONICET, Buenos Aires, Argentina.*

[§]*Department of Microbiology, New York University, Alexandria Center for Life Sciences, New York, United States.*

[†]*Protein Physiology Laboratory, Instituto de Química Biológica de la Facultad de Ciencias Exactas y Naturales (IQUIBICEN) CONICET- Universidad de Buenos Aires, Argentina.*

Running title*: Hidden structural codes in protein intrinsic disorder*

*To whom correspondence should be addressed: Gonzalo Prat Gay, *Protein Structure-Function and Engineering Laboratory, Fundación Instituto Leloir and Instituto de Investigaciones Bioquímicas de Buenos Aires (IIBBA) CONICET. Av. Patricias Argentinas 435 1405 CABA, Argentina. Telephone: (0054) 11523875000; FAX: (0054) 1152387501; E-mail:* gpg@leloir.org.ar

ABSTRACT

Intrinsic disorder is a major structural category in biology, accounting for over 30% of coding regions across life domains, yet it consists of conformational ensembles in equilibrium, a major challenge in protein chemistry. Anciently evolved papillomavirus genomes constitute an unparalleled case for sequence to structure-function correlation in cases where there are no folded structures. E7, the major transforming oncoprotein of human papillomaviruses is a paradigmatic example among intrinsically disordered proteins. Analysis of a large number of sequences of the same viral protein allowed for the identification of a handful of residues with absolute conservation, scattered along the sequence of its N-terminal intrinsically disordered domain, which intriguingly are mostly leucine residues. Mutation of these led to a pronounced increase in both α-helix and ß-sheet structural content, reflected by drastic effects on equilibrium propensities and oligomerization kinetics, and uncovers the existence of local structural elements that oppose to canonical folding. These folding relays suggest the existence of yet undefined hidden structural codes behind intrinsic disorder in this model protein. Thus, evolution pinpoints conformational hotspots that could have not been identified by direct experimental methods for analyzing or perturbing the equilibrium of an intrinisic disordered protein ensemble.

Keywords: *Intrinsically disordered proteins, E7, Virus, Papillomavirus, Protein evolution, Protein folding, Mutagenesis*

Proteins require structural codes that allow chemistry to operate biological function. These codes are no longer recognized as rigid, unique or even rendering defined structures represented by a set of atomic coordinates[1]. Intrinsically disordered proteins (IDPs) and domains (IDDs) are unquestionably a major structural category among proteins. They were initially defined experimentally as natively unfolded proteins[1] but found to represent roughly over 30% of coding regions[2,3]. IDPs are frequent among intracellular signaling networks and processes that involve protein interactions in both physiology and pathology across living organisms[4-6]. IDPs were initially defined as the absence of folded structure and the binary concept of disorder and order are now recognized as coexistent within a protein, and a convergence of techniques with NMR as the most powerful and promising are starting to provide dynamic and realistic quantitative representations of conformational ensembles of proteins[7].

IDPs are particularly overrepresented in virus genomes where the economy of functional versatility evolved for minimal genetic information[8]. Eukaryotic viruses frequently use a number of gene products that shift the cellular activity in their favor by interfering with cellular pathways. Human papillomaviruses (HPVs) are the etiological agent for cervical cancer, among other malignancies[9], and its tumorigenic activity is linked to at least two oncoproteins, E6 and E7, which are paradigms for hijacking the cell cycle[10]. Intrinsic disorder plays a critical role in E7[11], the main cellular transforming factor which exerts its effect by targeting the retinoblastoma (Rb) tumor suppressor for proteasomal degradation[10,12]. This causes the release of the general transcription factor E2F responsible for the transcription of a set of genes required for DNA synthesis, indispensable for genomic replication of a virus with no such machinery[10].

HPV-16 E7 is a 98 amino acid acidic protein with a highly promiscuous target binding activity despite its small size[11,13-15]. It consists of a tetrahedral zinc stabilized globular C-terminal domain (E7C, residues 51-98) and a N-terminal IDD (E7N, residues 1-40), which contains the binding motif of the primary target Rb plus several other linear motifs, linked by a proline-rich hinge region (E7H) (Fig. 1)[11,13,14]. We had earlier determined E7's intrinsically disordered nature and described it as an extended plastic dimer, with twice its expected hydrodynamic volume if globular[16]. We experimentally defined E7N as a *bona fide* domain, despite its small size, the lack of canonical secondary or tertiary structure, and cooperative unfolding[13,17]. Structures were determined for the E7C of HPV-1A and HPV-45 types, but these could only be attained in the absence of the disordered E7N domain[18,19]. No structure for HPV-16 is available despite its high medical relevance. We have shown that HPV-16 E7 can also self-assemble into defined spherical oligomers (E7SOs) with amyloid-like properties and chaperone holdase activity upon removal of its coordinated zinc atoms, where the linear motif exposing E7N IDD faces the solvent, providing solubility to the otherwise aggregation prone E7C oligomer[11,13,20-23].

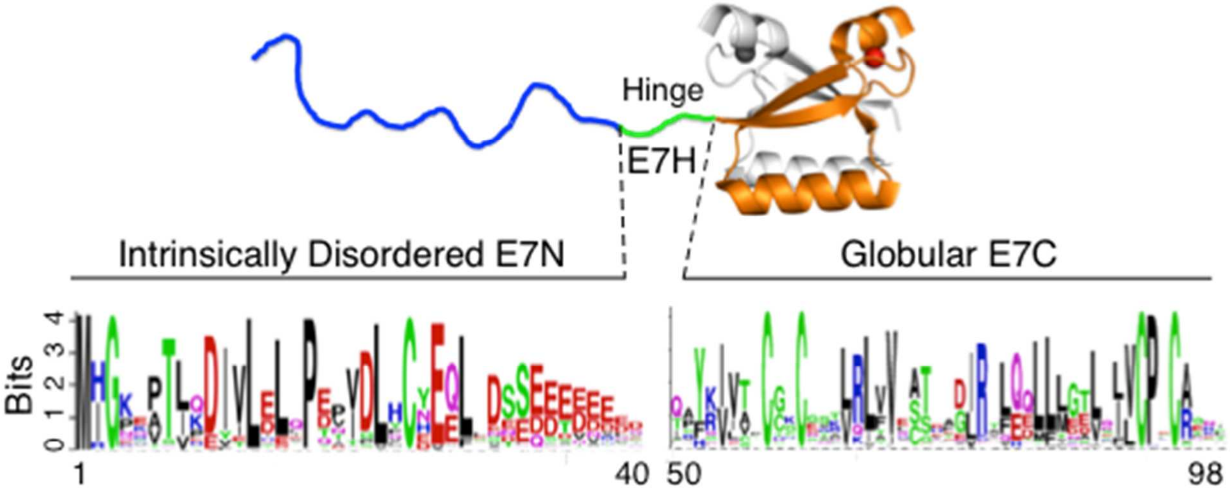


Figure 1. ***Schematic representation of the structural and sequence properties of the HPV-16 E7 protein*** *(Upper panel) Domain organization of the E7 protein. The intrinsically disordered N-terminal domain E7N (blue) is joined to the globular dimerization domain E7C (orange) by a flexible hinge region E7H (green). For clarity, only one monomer of the E7 dimer is depicted, with the second E7C unit in gray (PDB ID: 2F8B). (Lower panel) Sequence logos of the E7N (left) and E7C (right) domains obtained from the alignment of 137 sequences of alpha and beta HPV species E7 proteins. Numbering corresponds to the HPV-16 E7 protein. The height of the letters in each position is correlated to conservation of the respective residue.*

Papillomaviruses are an ideal family of viruses for sequence evolutionary studies since there are a few hundreds of anciently evolved and genetically stable variants (types) that provide a rich substrate for addressing sequence conservation in connection with structure and function[24]. We have found that E7N and E7C show similar degrees of conservation and co-evolution of particular residues indicative of interactions that could only be inferred by sequence analysis[14,24]. However, an imperative question arises related to the presence of a discrete number of hyperconserved residues (more than 80% or 4 bits Shannon entropy) in a region that bears no defined folded structure. An additional challenge is how to probe structure in a dynamic ensemble[25], and non-classical approaches are required to investigate IDP structrure, such as the combination of conformational sampling and NMR parameters[26]. Nevertheless, identification and mutation of these conserved residues constitutes a powerful, almost unique, approach.

We show that the highly conserved residues do not respond to RNA sequence evolution events related to known translation, transcription or splicing, but to yet unknown structure-folding reasons. Single mutation of these residues leads to increase of structural content in the different conditions that were purposely instrumented to test the ensemble. This increment is independent of the type of secondary structure, where the conserved residues participate as local structural elements. Mutations of these elements release structural transitions leading to the conformational diversity that can be adopted by the intrinsically disordered ensemble (IDE) of E7. In this work we show that sequence evolution guided-mutagenesis is a unique source of perturbation of IDEs, which hopefully, acting in complementation with advanced NMR and direct techniques, will provide integral and dynamic representation of the structural basis of intrinsic disorder.

MATERIALS AND METHODS

Sequence and nucleotide conservation analysis – We used 137 protein sequences corresponding to all alpha and beta species E7 proteins[14]. We collected a single non-redundant nucleotide sequence corresponding to each protein UNIPROT ID using the Uniprot database (http://www.uniprot.org/uploadlists/). We then used the tool TranslatorX[27] (http://www.translatorx.co.uk/) to obtain a nucleotide alignment by back-aligning the nucleotide dataset to the previously obtained protein dataset alignment, and produced nucleotide and aminoacid sequence logos for each alignment. In order to search for conserved RNA structures, we used the RNAZ algorithm[28] which uses an alignment of up to six nucleotide sequences to search for conservation of RNA secondary structure elements. We used an alignment of E7 nucleotide sequences from six types within the alpha 9 HPV species (HPV-16, HPV-31, HPV-33, HPV-35H, HPV-52, and HPV-58), which includes the HPV-16 type. Considering that, even within this closely related phylogenetic group, no conserved RNA structures were detected, the search did not expand to more divergent alignments.

Expression and purification of proteins – HPV-16 E7 wild type (E7wt) and E7Δ1-26 (a truncated variant with deletion of aminoacids 1-26) were cloned as thrombin cleavable proteins to the maltose binding protein (MBP) into a pMALp2 vector (New England Biolabs, Beverly, MA), expressed in E. coli TB1 strain and purified as previously described[16,29]. E7Δ1-26 was used as control, being representative of E7C, due to the difficulty in obtaining E7C for its tendency to aggregate. Using a site-directed mutagenesis strategy, constructs containing a sequence codifying for E7 with point mutations on selected conserved residues within the E7N were generated, amplified by inverse PCR and cloned into a pTZ18U-based vector (Pharmacia, Uppsala, Sweeden) under control of the T7 promoter downstream to a short peptide (19 aminoacids) of a β-galactosidase protein. The resulting expression vectors were transformed into BL21(DE3) Plyss E. coli strain and proteins were purified as previously described[17,30]. The following mutants were obtained: L13A (leucine 13 was substituted by alanine), L13G (Leucine 13 was substituted by glycine), L15G (Leucine 15 was substituted by glycine), P17A (proline 17 was substituted by alanine), L22G (leucine 22 was substituted by glycine), L28G (leucine 18 was substituted by glycine). All proteins were stocked in 10 mM sodium phosphate buffer pH 7.0 and 1 mM dithiothreitol (DTT). Additionally, E7N(1-40) a peptide comprising the entire E7N domain was synthesized by F-moc chemistry (W.M Keck Facility, Yale University, New Heaven, CT) and used as representative control of E7N. Protein concentration was determined by three different methods: Bradford assay, reverse phase HPLC and size exclusion chromatography (SEC).

Size exclusion chromatography (SEC) – SEC experiments were carried out using a Superdex-200 (Pharmacia Biotech, Uppsala, Sweeden) equilibrated with 12.5 mM sodium phosphate buffer pH 7.3 and 1 mM DTT. The elution of proteins was monitored at 220 nm. The void volume (V_0) and total volume (V_t) were determined by loading blue dextran and methionine, respectively.

Far-UV Circular dicroism (CD) – Far-UV CD experiments were carried out on a Jasco J-810 spectropolarimeter (Jasco, Japan), with cell paths of 0.1 cm, employing a scanning speed of 50 nm/min, a band width of 0.4 nm and an average response time of 4 sec. All spectra were an average of at least 8 scans. Baseline measurements using buffer alone were subtracted from the measured spectra. The temperature was set at 25 °C using a *Peltier* temperature-controlled sample compartment. Samples with protein concentration of 10 uM and peptide concentration of 20 uM were used for the experiments. For pH dependence experiments, an initial measurement in 10 mM sodium phosphate buffer pH 7.5 was followed by a second measurement after addition of 100 mM sodium formiate buffer pH 3.0. Raw data were converted to molar ellipticity [θ] using the following equation[31]:

$$[\theta] = \frac{\deg}{[c]*\#\,bonds*L*10{,}000} \qquad (1)$$

where deg is the raw signal in millidegs, [c] is protein concentration in molar units, # *bonds* is the number of peptide bonds (number of aminoacids - 1), and L is the path length in cm.

Analysis of α-helix content by 2,2,2-trifluoroethanol (TFE) measurements – TFE stabilization was carried out at 25 °C. Samples were dissolved in 10 mM Tris HCl buffer pH 7.5 and equilibrated in 0 - 53 % TFE (volume of TFE added/total volume added) as previously described[29]. The mean residue ellipticity at 220 nm, which is assumed to be proportional to helical content, was plotted as a function of [TFE]/[H_2O] ratio and was fit to a two-state ¨coil to helix¨ equilibrium model[32]. This model assumes that the free energy for α-helix formation depends linearly on [TFE] / [H_2O] ratio. Data were fitted using the following equation:

$$[\theta] = \frac{[\theta]^{TFE} + [\theta]^{H_2O} * \exp\left(\left(-\Delta G^{H_20} - m[H_2O]/[TFE]\right)/RT\right)}{1 + \exp\left(\left(-\Delta G^{H_2O} - m[H_20]/[TFE]\right)/RT\right)} \qquad (2)$$

where $[\theta]^{H2O}$ and $[\theta]^{TFE}$ are the mean residue ellipticities in water and at high TFE concentration. R is the gas constant and T is the temperature in kelvin. The thermodynamic parameters ΔG^{H20} and m value were obtained from the data fit.

The average percentage of residues in α-helix conformation in water and TFE for each protein was calculated from the values $[\theta]^{H2O}$ and $[\theta]^{TFE}$, using the following empirical equation to define molar ellipticity value expected for 100% of α-helix[31]:

$$[\theta] - 39{,}500 * \left(1 - \frac{2.57}{n}\right) \qquad (3)$$

where n is the number of residues.

Aggregation kinetic measurements – Aggregation kinetics were recorded in a Jasco V-550 UV spectrophotometer (Jasco, Japan) by following scattering signals on 10 uM of E7wt and mutants in 10 mM of sodium phosphate buffer pH 7.5 during 5 minutes. After this initial measurement, the sample pH was decreased by adding 100 mM of sodium formiate buffer pH 3.0. Immediately after mixing, scattering was recorded until aggregation traces reached saturation. The temperature was controlled by a thermal bath fixed at 25 °C. Considering that in all mutants aggregation was rapid, this experiment was repeated using protein samples at a concentration of 2.5 uM in order to obtain kinetic traces more suitable for data fitting. The kinetic traces for E7wt, L15G and L28G were fitted to the following empirical sigmoid function[33]:

$$[Absorbance]360nm = A_0 + \frac{A}{1 + \exp\left[\left(t_{1/2} - t\right)\right]k_{agg}} \qquad (4)$$

where A_0 is mean residue absorbance at time 0, A is the total change of mean residue absorbance from time 0 to ∞, $t_{1/2}$ is the middle point of aggregation, and k_{agg} is the apparent aggregation rate constant.

Considering that in most of the mutant proteins aggregation was rapid and the *lag* phase was not measurable, the following single exponential function[33] was used for fitting the kinetic traces of the rest of the mutant proteins:

$$[Absorbance]_{360nm} = A\exp\left(-k_{agg} * t\right) + q \qquad (5)$$

where A and k_{agg} have the same meaning as for eq. #4, and q is the final mean residue absorbance at time ∞. For better comparison of the k_{agg} values, all the kinetic traces, including those of proteins that exhibited *lag* phase (E7wt, L15G and L28G), were fitted using eq. #5, selecting in these cases only the data points of the exponential phase.

Far-UV CD kinetics were measured using 10 uM of protein samples at a fixed wavelength of 205 nm (at which a notorious difference in the CD spectra was observed between pH 7.5 and pH 3.0), following the same conditions described for UV scattering (an initial measurement at pH 7.5 for 5 minutes, followed by addition of 100 mM of sodium formiate buffer pH 3.0 and a kinetic measurement at this condition for 25 minutes). The data obtained from the CD aggregation kinetics concurred with those obtained from scattering. Nevertheless, the CD data was not suitable for fitting.

Dynamic light scattering (DLS) – DLS measurements were taken on a Zetasizer Nano S DLS device (Malvern Instruments). Measurements were carried out on 50 uM of protein sample in 10 mM sodium phosphate buffer pH 7.5 at 25 °C, followed by addition of 100 mM sodium formiate buffer pH 3.0. Under the latter condition, measurements were taken at time-points 0 and 30 minutes. Prior to the experiments protein samples and buffers were filtered with Ultrafree-MC microcentrifuge filters (0.22 uM, Millipore).

RESULTS

Origins for strong residue conservation in the E7N IDD – The papillomavirus E7 protein is composed of an intrinsically disordered N-terminal domain (E7N IDD)[17] evolved in close proximity to a globular C-terminal dimerization domain (E7C), connected by a flexible linker or "hinge" (E7H) (Fig. 1)[14]. Based on an extensive set of 200 sequences spanning a wide evolutionary range (~350 Mya), we showed that despite its intrinsically disordered nature, the E7N domain displays similar levels of sequence conservation to those of the globular E7C domain. This is even more evident if the zinc coordination cysteins are not considered, leaving side chains that can only be involved in non-covalent and non-metal folding interactions[14]. While sequence conservation within E7C can be explained by conservation of the globular fold, conservation within E7N can be only partially explained by the presence of linear motifs that span short local stretches of sequence (Fig. 2A)[14]. Surprisingly, strong sequence conservation is observed in positions scattered along the sequence of E7N that do not correspond to functional motifs, suggesting that yet uncovered conformational features could be conserved within this domain[14]. The length of the linker (E7H) is also rather conserved, mostly between 5 and 15 residues (Fig. S1). In addition to its conserved short length, E7H is rich in proline residues, which isomerization was shown to govern an extremely slow antibody-antigen binding reaction[34]. This is a strong indication that it plays a role in connecting both domains with a conserved proximity, with so far undefined long range interdomain non-covalent transient contacts between E7N and E7C.

Why are there so many highly conserved residues in an intrinsically disordered domain? First, we sought to address whether this conservation originates at the amino acid or nucleotide levels. We analyzed codon conservation for each residue within the E7N alignment and observed that nucleotide conservation in all positions followed mainly the frequency distribution expected for all codons (Fig. 2A), strongly suggesting that nucleotide variation was equilibrated within the time period analyzed, and provided no evidence for selection pressures acting at the nucleotide level. We also analyzed the E7N nucleotide sequences in search for conservation of known regulatory sites, such as splice acceptor and donor sites[35], presence of CpG islands, and internal late promoters such as p670 described for HPV-16 E7[36], and we found no correlation with the highly conserved positions (Not shown). Finally, we found no evidence for RNA structure conservation as a plausible reason (Fig. S2). Altogether, these observations point at the conservation of residues responding to structure-folding events.

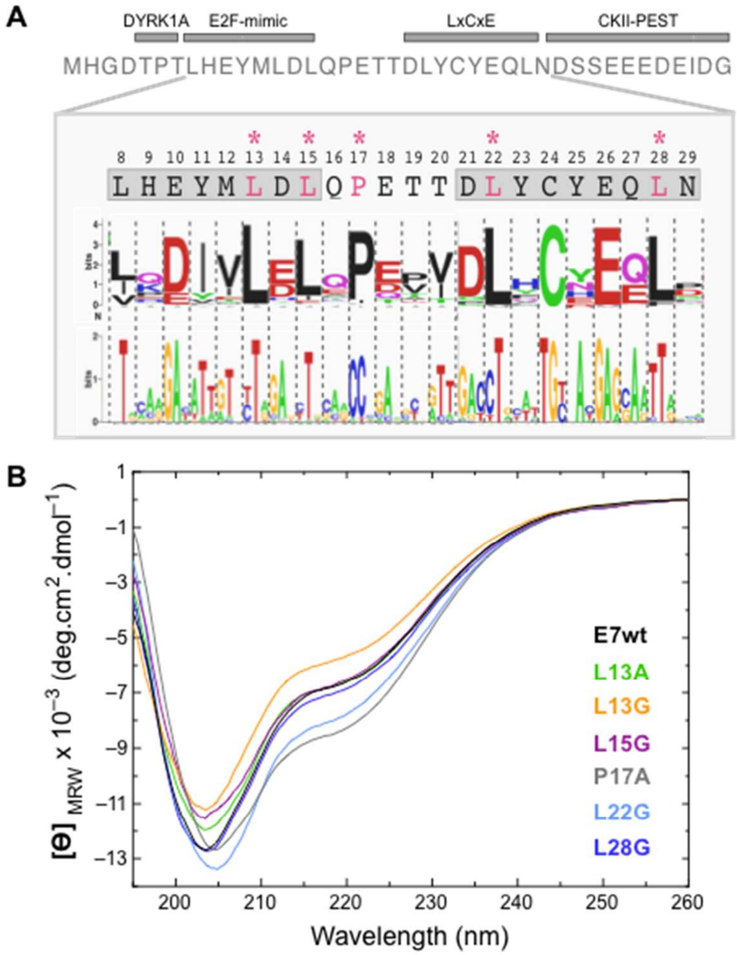


Figure 2. ***Sequence conservation of aminoacids and nucleotides within HPV-16 E7N, and secondary structure of E7wt and E7 IDD mutants***. *(**A**) (Upper panel) Sequence of the HPV-16 E7N domain (residues 1-40) and positions of known functional elements within the sequence. (Lower panel) Aminoacid and nucleotide sequence logos of the region encompassing residues 8-29. Positions mutated in this study are marked in pink and highlighted by an asterisk. Linear motifs are shown as gray boxes. Nucleotide positions are aligned below the aminoacid sequence, dashed lines separate individual codons. (**B**) Far-UV CD spectra of E7wt and E7 IDD mutants in 10 mM sodium phosphate buffer pH 7.5 at 25 °C. E7wt is shown in black, mutants are represented in colors.*

Mutation of highly conserved residues in E7N IDD increases α-helix content – We had previously determined by nuclear magnetic resonance (NMR) that the isolated E7N shows random coil-like chemical shifts, where 2,2,2-trifluoroethanol (TFE) can stabilize two pre-existing α-helices spanning residues L8-L13 (Helix I) and P17-N29 (Helix II)[29]. There is no structure for the full length HPV-16 E7 or any other virus type, since the dynamics of the E7N perturbs the spectra[18,19], something evident from inspection of the HSQC spectrum (Fig. S3). Therefore, we used far-UV circular dicroism (CD) spectra as a robust and highly sensitive method for assessing secondary structure. At pH 7.5, HPV-16 E7 wild-type (E7wt) showed the reported typical spectrum of a mixed folded/disordered conformation, with a minimum at 204 nm and a band and 222 nm (Fig. 2B and Fig. S5)[16]. On the other hand, the spectrum of E7N IDD is characteristic of a disordered polypeptide, with a minimum around 200 nm[17,29]. E7Δ1-26, a variant lacking the first 26 residues, showed a loss of signal at the 204 nm minimum under the same condition (pH 7.5), in comparison with full-length wild-type E7 (Fig. 4A and Fig. S5), indicative of a lower disorder/order structural ratio. Differential spectra showed that E7N lacks stable canonical secondary structure and behaves as an IDD also in the context of the full-length protein[17]. In a previous work, we showed that E7N populates different conformational species at different solvent conditions with their own characteristic CD spectra[17], and this provides us with a valuable tool to probe the effect of mutation on the structure content within the IDD.

In order to address the role of the hyperconserved residues, we mutated the four most conserved leucine residues (positions 13, 15, 22 and 28) and a proline (position 17) (Fig. 2A). We mutated the leucines to glycine instead of alanine, since we hypothesize them to be involved in non-polar interactions and we intended to prevent any possible side chain interactions, and the proline residue was mutated to alanine. Mutants showed the same anomalously extended hydrodynamic behavior of the wild-type protein previously described (Fig. S4)[16]. The secondary structure of full-length E7 turned out to be very sensitive to the single mutations in the E7 IDD at neutral pH, as judged by modifications in the CD spectra corresponding to a conformational ensemble (Fig. 2B and Fig. S5). The most salient feature was the increase in α-helical content as indicated by the increase in the negative minimum at 222 nm for P17A and L22G, and the opposite effect for L13G (Fig. 2B and Fig. S5). Although not affecting the 222 nm minimum, L15G showed a change in the ratio 204/222 nm which indicates a significant change in the structure population of the IDE (Fig. 2B and Fig.S5). A more accurate quantitative analysis of the effect of the mutations on α-helix stabilization of the IDE can be obtained from TFE titrations[29,32]. Gradual addition of TFE led to an increase in α-helix content judged by the unequivocal minima of the CD spectra at 208 and 220 nm (Fig. 3A and Fig. 3B). Mutations had a marked effect on helix stabilization, with the large increase observed in L28G represented as an example (Fig. 3A, inset). We carried out TFE titrations of all the variants and based on the known formalism of the effect of TFE on α-helix equilibria[32], we calculated the percentage of α-helix content from extrapolation of the titrations, at zero or high TFE concentrations (Fig. 3C, Table 1). Extrapolation of the wild-type data to buffer yielded a 18.6 ± 0.4 % of α-helix (α_i, 0% TFE) (Table 1) in excellent agreement to what is expected from the structures of E7C (16-17%)[18,19]. Similarly, extrapolation of ellipticity values from high TFE, yielded the α-helix content at solvent saturation (α_f, 53% TFE), which in the case of wild-type is 32.9 ± 0.9 % (Table 1). For a simplified evaluation of the effect of each mutant, we subtracted the value from the wild-type and compared effects on α_i and α_f (Fig. 3C, Table 1). All mutants showed an increase in α-helix content, indicative of a release of structure caused by each replacement (Fig. 3C, Table 1), except for L13G that showed no effect. This is in apparent contradiction with a decrease in α-helix in buffer (Fig. 2B), but we consider the quantitative analysis from the extrapolation is more accurate. An additional major conclusion is that the effect is different in water than in high TFE, with the smallest difference in α_f -α_i in L13A and the largest in L28G (Fig. 3C, Table 1). It is safe to assume that the TFE α-helix is consolidated[29] differently from the pre-helical or nascent helix-like structures that may take place in water, and the differential effect of mutations is precisely reflecting this fact. It is thus tempting to suggest that some of the mutations, coincident with boundaries of the two α-helices as determined by NMR[29], merely follow helix propensity at N- or C- caps. This could be the case of P17A mutation, where a helix-breaking was replaced by a helix-stabilizing residue. However, glycine had the lowest helix propensity after proline, and leucine had the second highest helix propensity after alanine[37], yet the absolute largest increase in α-helix was observed in L28G, and this applies to L22G and L15G.

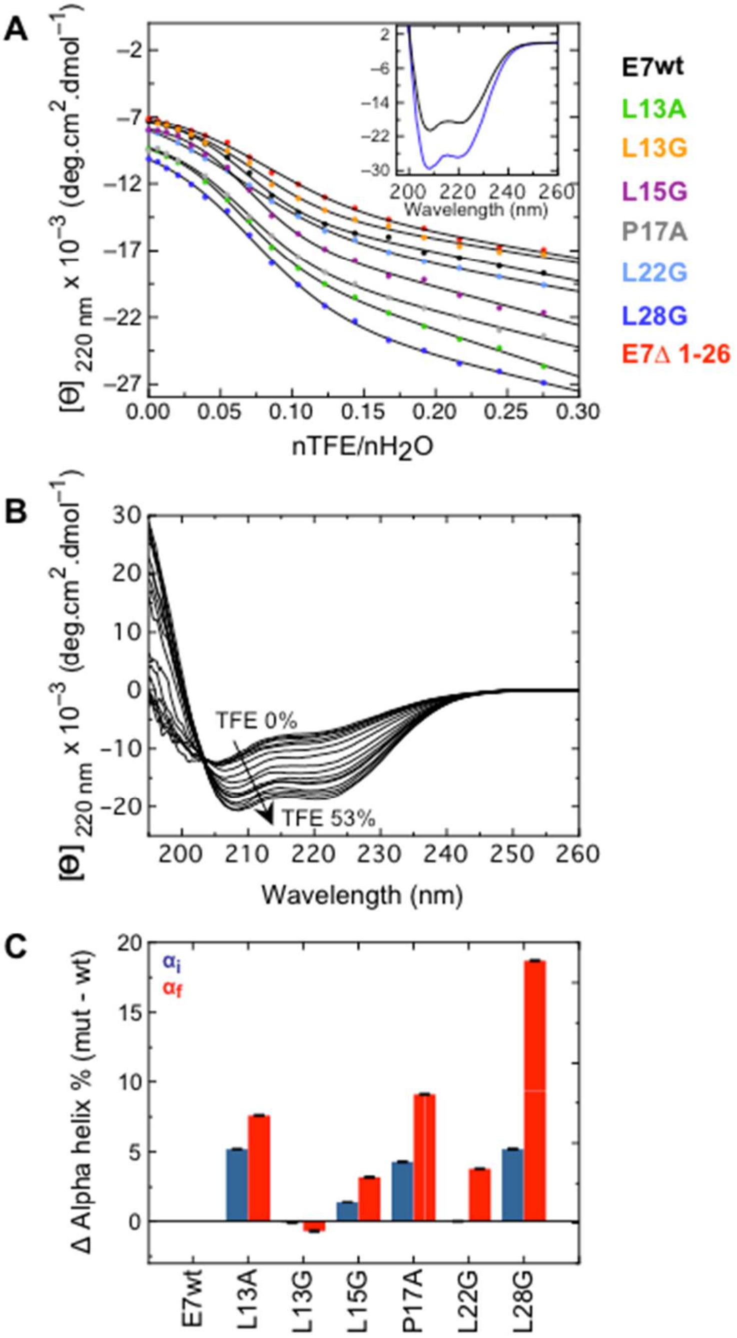


Figure 3. ***Characterization of α-helical populations within HPV-16 E7 and E7 IDD mutants by Far-UV CD spectroscopy****. (A) Titration curves for E7wt and mutants following ellipticity at 220 nm as a function of [TFE]/[buffer] molar ratio. 10 mM Tris HCl buffer pH 7.5 was used for the experiment. The temperature was set at 25 °C. Black circles represent E7wt, color circles are used to represent mutant proteins. Solid lines show fitting of the data to a two-state "coil-helix" equilibrium model obtained according to eq. #2. Inset: Far-UV CD spectra of E7wt (black) and L28G (blue) in 10 mM Tris HCl buffer pH 7.5 containing 53 % TFE. (**B**) Far-UV CD spectra of E7wt in 10 mM buffer Tris HCl pH 7.5 at 25 °C and TFE (v/v) percentages ranging from 0 % to 53%. The arrow indicates the sense of change upon increasing TFE. (**C**) Initial (blue bars) and final (red bars) percentages of α-helix contents induced by TFE in E7 mutants relative to E7wt. The percentages of residues in α-helix conformation were calculated from the data fit mentioned above.*

The stability of the α-helix formed is determined by the free energy obtained from the titration experiments, extrapolated to water (ΔG^{H20}) (Table 1). These free energies were positive because energy is provided by the addition of the solvent, the smaller the value the higher the stability. They are in the range observed for peptides and fragments, where no other energetic components from tertiary interactions in a cooperative fold should be present and reflected in overall larger folding free energies. Interestingly, the largest stabilization was caused by L22G and L28G mutants (Table 1), which appeared to follow the same rationality of a higher stability caused by a residue with helix propensity at the opposed extreme of the one replaced. Altogether, these results strongly suggest that the mutation does not generate a more stable helix because of stabilization of the α-helix in TFE *per se* (the "end state"), but mutation of the leucine residues is destabilizing a local yet undefined structure opposing to helix formation nucleated around a leucine residue, and the effect of this is a shift of the equilibrium towards a folded helix. In the case of P17A, the effect may well be that expected from removing the helix breaking proline and the consequent stabilization or extension of a α-helix in TFE. The change in *m* values was particularly noticeable in L22G and L28G (Table 1), which are within the α-helix II in TFE. Our interpretation is that a drastic change in *m* value responds to a change in the structure at either or both sides of the equilibrium[38]. Since we know that α-helices I and II are indeed formed at high TFE[29], we propose that this drastic change reflects a different structure in the IDE, where α-helix I is more coil-like while α-helix II, containing L22G and L28G involve stable structures in the form of nuclei that oppose to helix formation.

Table 1. Parameters for α-helix stabilization in HPV-16 E7 protein and E7 IDD mutants

	ΔG^{H2O}	m	% α-helix[a]		
	(Kcal/mol)	(Kcal mol^{-1} (% TFE)$^{-1}$)	$H_2$0 (Initial, α_i)	TFE (Endpoint, α_f)[b]	α_f - α_i
E7$_{wt}$	2.0 ± 0.2	28.2 ± 3.1	18.6 ± 0.4	32.9 ± 0.9	14.4
L13A	2.0 ± 0.4	28.0 ± 3.9	23.8 ± 0.4	40.5 ± 0.9	16.7
L13G	1.8 ± 0.3	23.7 ± 3.8	18.5 ± 0.7	32.3 ± 1.5	13.8
L15G	1.9 ± 0.2	28.2 ± 3.1	20.0 ± 0.5	36.1 ± 1.2	16.1
P17A	1.5 ± 0.1	22.8 ± 2.0	22.9 ± 0.6	42.0 ± 1.1	19.1
L22G	1.1 ± 0.1	17.1 ± 1.1	18.6 ± 0.5	36.7 ± 0.8	18.1
L28G	1.3 ± 0.1	18.5 ± 1.5	23.8 ± 0.8	51.7 ± 1.7	27.9
E7Δ1-26	1.6 ± 0.2	18.6 ± 2.8	17.8 ± 0.6	31.2 ± 1.9	13.4

[a] Percentage of residues in α-helix conformation

[b] Values calculated from molar ellipticity (220 nm) obtained from fitting with eq. #2

ß-sheet aggregation is drastically affected by mutation of conserved residues – E7 proteins are strongly acidic, with pKas around 4.5. In order to uncover and probe further possible conformational biased populations, we analyzed E7 in pH conditions where negative charges are neutralized. At pH 3.0, there was a complete disappearance of the ca 204 nm disordered component present at pH 7.5 (Fig. 4A). Although the resulting spectrum corresponds to an increase in ß-sheet structure (Fig. 4A), the minima at around 220 nm also suggest contribution of α- helix, most likely remaining native structure in globular E7C. The truncated E7Δ1-26 variant showed an evident α-helix rich spectrum at pH 7.5, devoid of the disordered contribution by the first 26 residues (Fig. 4A), strongly suggesting that the increase in ß-sheet requires the entire E7N IDD and occurs at the expense of the loss of disorder (Fig. 4A). There was a further increase in α-helix of E7Δ1-26 at pH 3.0, as the shift and increase of minima at ca 208 nm and 220 nm indicate (Fig. 4A). The ß-sheet enriched full-length E7 species formed a metastable monodisperse oligomer of 100 nm size and a large polydisperse aggregate, while the α-helical E7Δ1-26 species at pH 3.0 formed a monodisperse oligomer of 20 nm and no aggregate, as determined by dynamic light scattering (DLS) (Fig. 4A, inset). Under similar conditions, there was no oligomer or aggregate formed in the E7N(1-40) fragment (Not shown). These two results indicate that not only the full sequence information of the E7N IDD is required for ß-sheet oligomerization and aggregation, but it also requires both domains of the protein, strongly suggesting that the first 26 residues, in the context of the full-length protein, are required for ß-sheet formation.

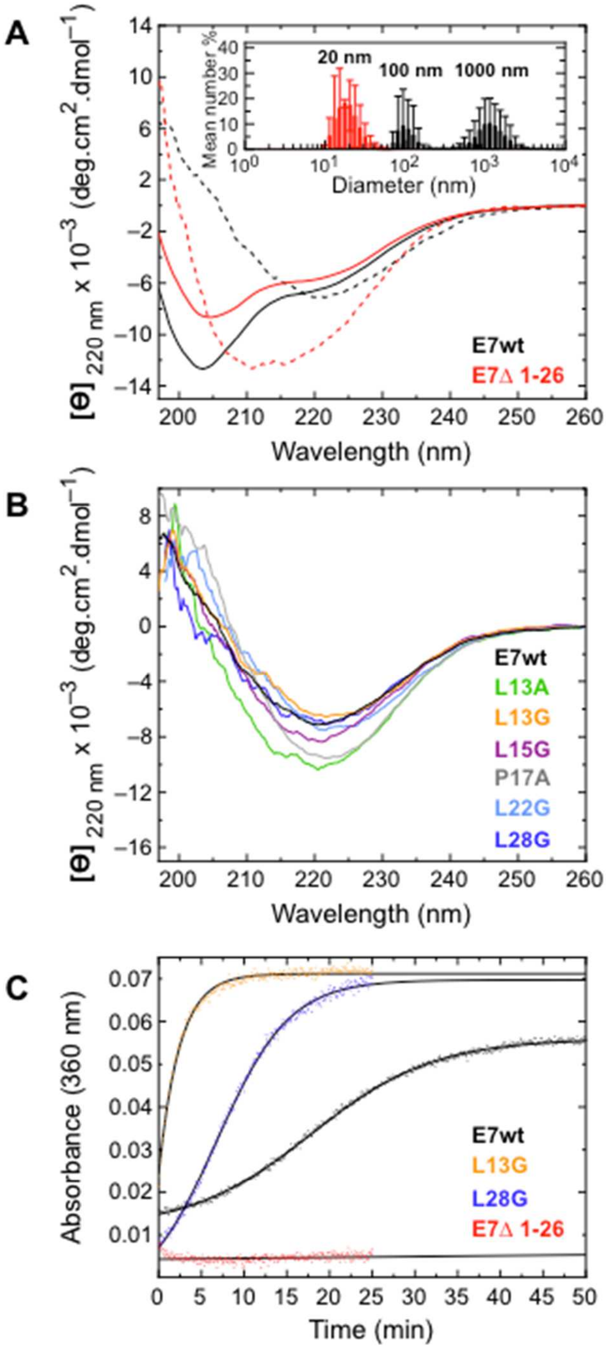


Figure 4. ***pH dependence oligomerization of HPV-16 E7 is governed by E7N.*** *(A) Far-UV CD spectra of E7wt (black) and E7Δ1-26 (red) in 10 mM sodium phosphate buffer at pH 7.5 (full line), followed by addition of 100 mM sodium formiate buffer at pH 3.0 (broken line). Inset: Size distribution measured by DLS at pH 3.0. Measurements were performed in E7wt (black) and E7Δ1-26 (red) immediately after decreasing the sample pH (7.5) by addition of 100 mM sodium formiate buffer pH 3.0. (B) Far-UV CD spectra of E7 and mutants, followed by acidification of the environment by addition of sodium formiate buffer pH 3.0 to samples previously measured at pH 7.5. (C) Aggregation kinetics determined by scattering at 360 nm. Comparison of kinetic traces of E7wt (black) and two representative E7 mutants, L13G (yellow) and L28G (blue). E7Δ1-26 (red) did not aggregate under this condition. Solid lines show the best-fit profiles obtained with sigmoid or exponential functions (eq. #4 and eq. #5).*

All mutants showed an overall similar behavior at pH 3.0 but with different qualitative and quantitative degrees of ß-sheet enrichment, with L13A, L15G and P17A being the most salient (Fig. 4B and Fig. S6A), even though the structural details were not possible to infer at this time. However, aggregation kinetics was highly sensitive to mutation, under this acidic environment where, surprisingly, all the mutations increased the rate (Fig. 4C and Fig. S6B). The wild-type E7 showed a *lag* phase where a nucleation process most likely takes place, and this tends to disappear as the aggregation rate of the mutants increases. We propose that mutation stabilizes the aggregation nucleus, which is defined as an aggregation competent low molecular species, often monomer or dimer[39]. This species is partly folded, as is the case in most ß-oligomerization processes, where a certain degree of folding, but not complete or compact structure, is required for the self assembly to take place[39]. Mutation stabilizes the aggregation nucleus and thus lowers the kinetic barrier because it destabilizes a folded structure, which population in the IDE is increased under this acidic condition. Except for the most moderate effect of L28G, the rest of the mutants (positions 13 to 22) accelerated the rate between 2.7 and 3.5 fold (Table 2). E7N(1-40) did not spontaneously form ß-sheet structure or aggregates at pH 3.0 (Not shown), and we know that residues 1-26 are required for ß-structure formation, since E7 Δ1-26 could neither increase its ß-sheet content nor aggregate (Fig. 4A, Fig. 4C, and Fig. S6B).

Table 2. Parameters for aggregation in HPV-16 E7 protein and E7 IDD mutants

	k_{agg}	t_{lag}	$t_{1/2}$	k_{agg} mut/k_{agg} E7$_{wt}$
E7$_{wt}$*	0.13 ± 0.001	2.5	17.3 ± 0.1	1.0
E7$_{wt}$	0.08 ± 0.001		14.7	1.0
L13A	0.28 ± 0.003		2.3	3.6
L13G	0.30 ± 0.007		0.7	3.8
L15G	0.21 ± 0.003		2.3	2.7
P17A	0.27 ± 0.002		1.3	3.5
L22G	0.22 ± 0.002		2.6	2.9
L28G	0.12 ± 0.002		7.4	1.5

* Parameters obtained using a sigmoid function

k_{agg}, apparent aggregation rate constant

t_{lag}, lag time

$t_{1/2}$, apparent aggregation middle time

DISCUSSION

The sequence-structure paradigm determines that the information required for a polypeptide to fold into a discrete structure is provided by the amino acid sequence[40]. IDPs can be described as molecules with no defined or compact structure[1], in which it is assumed that there are not enough local and long-range native interactions to stabilize a particular fold, as in globular proteins. The question is whether this lack of sequence information for a compactly folded structure is the single prerequisite for a protein to be intrinsically disordered.

An additional far from trivial challenge is the fact that IDPs are conformational ensembles consisting of a mix of populations, which are believed to be discrete yet dynamic and in fast exchange[1,41,42], not amenable for most structural analysis methods of atomic detail. Moreover, we know for a fact that there will be no atomic details, no fixed or defined structures, and that is the essence of IDPs in biology. To experimentally overcome this, we established conditions where alternative α-helix or ß-sheet enriched conformers of the ensemble were stabilized, and the effect of the mutations was quantitatively evaluated. Mutations are ideal to probe structure and function, and are fairly easy to select for rational perturbation every time there is an atomic structure available[43]. This is not the case of IDPs, and this makes the papillomavirus evolution not only an unusual but also a unique tool for investigating "structure in the absence of structure" or in a mixture of conformers, by experiment[14,24]. Sequence analysis of HPV E7s pinpointed a handful of residues in the E7N IDD, which are highly conserved for reasons not related to nucleic acid sequence or function (Fig. 1), leaving yet unknown structural traits related to intrinsic disorder as the main explanation[14,24].

Several important conclusions can be obtained from the effect of the mutations on the α-helix population of E7N in the context of full-length E7. First, that mutations at all positions produced an increase in the structure content, and this effect was different in the consolidated α-helices in TFE (α_f) from the nascent structures in aqueous buffer (α_i), with no clear correlation in the α_f/α_i ratio of the mutants. This indicates that the structures involved are different, and the structure increment can be either stabilization of the consolidated (α_f) or the destabilization of a proto-structure (α_i) that shifts the equilibrium towards a consolidated α-helix. Although both showing a substantial increase in α-helix, L28G showed a much larger α_f/α_i ratio than L13A. The former had the largest effect on helix stabilization in TFE, despite glycine being a poor helix former and leucine being the second highest helix former after alanine. This strongly suggests that the stabilization of the α-helix is the result of a destabilization of structure nucleated around L28. Moreover, P17A attained only half of the α-helix increase of L28G, yet we replaced the best helix breaker with the best helix former, once again highlighting the presence of folding nucleation around L28. A similar line of reasoning, with effects of different magnitudes, can be applied for all four hyperconserved leucines. Leucine side chain is all apolar and buries 137 Å^2 of solvent accessible surface area, corresponding to a transfer free energy of 1.6 kcal.mol^{-1}[44,45]. In any case, the model we envision is that of individual local structure nucleation sites around each conserved leucine residues, not a structure based on long-range interactions of these residues.

It remains to be established, if possible, with specific highly sensitive NMR methods, what is the nature of the structure nucleation around each of these residues. This is rather challenging also because the assignment of the full length E7, both N- and C-terminal domains, has not been possible to date. What it is clear is that the combination of rich sequence information and mutagenesis in this unique system seems to be the only way we could learn about this phenomenon.

The formation of ß-sheet oligomeric structures at acidic pH took place at the expense of a loss of disorder at the E7N IDD, while the truncated variant E7(Δ1-26) did not increase the ß-sheet structure content, but showed a large increase in α-helix at the expense of the disordered contribution of residues 26 to 48 (Fig. 4A). These observations, together with the fact that E7N (1-40) did not lead to ß-oligomerization or aggregation, indicate that the full-length proteins, i.e. both domains, are required for the ß-sheet folding route. The large effect of mutations on the aggregation kinetics indicate that all positions play a role in ß-sheet stabilization by destabilizing structures that lead to a ß-oligomerization competent species, which we refer to as oligomerization nucleus. This is in line with the requirement of the entire E7N domain for stabilization of these aggregating prone ß-sheet conformers present in the IDE enriched at acidic pH. The fact that the mutants with the most increased ß-structure content in the metastable oligomer (L13A, L15G, and P17A) are the ones with the largest effect on the aggregation rate strongly suggests that both species are linked. Based on the cumulative results we propose that the limited ß-oligomerization responds to interactions formed within the E7N domain and between the E7N and E7C domains. On the other hand, given that E7 is a dimer, we propose that macroscopic aggregation responds to intermolecular interactions (Fig. 5). In any case, mutation of hyperconserved residues release and facilitate ß-sheet formation either by stabilizing endpoint ß-sheet or by destabilizing structure nuclei within E7N IDD that oppose to this folding route.

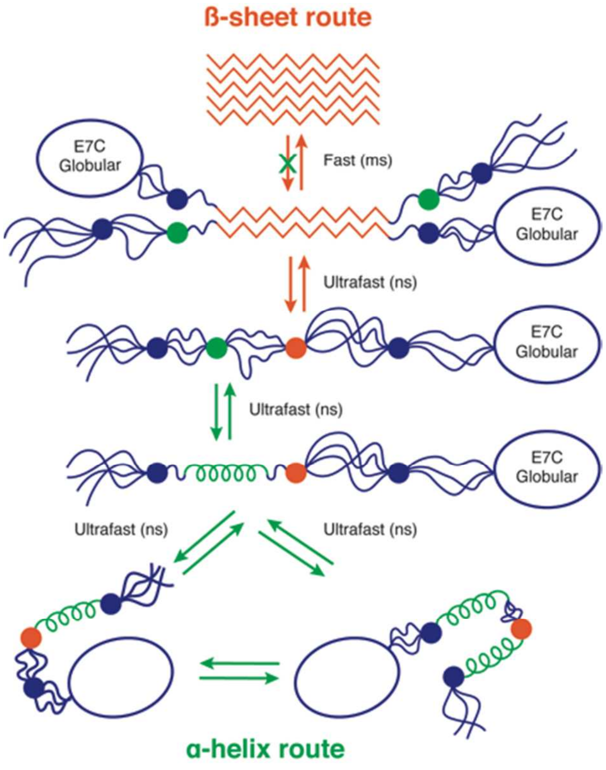


Figure 5. *Proposed model of the effect of single mutations on hyperconserved residues of E7 IDD.* *E7 monomer is shown as the start point of folding routes that can be followed after destabilization of anti-folding nuclei (knots). The β-sheet route is represented in red and the α-helix route in green. The remaining structure and anti-folding nuclei not involved in either β-sheet or α-helix routes are depicted in blue. The β-sheet route, triggered by acidic pH, is ultrafast, irreversible and at last leads to oligomerization and aggregation. On the other hand, the α-helix route is in fast exchange and reversible.*

The ability of certain sequences to form either α-helix or ß-sheet structures, named chameleon sequences, were shown to depend on the context within a protein[46], or cases were described in which α-helix and ß-sheet structure exchange in the millisecond time scale depending on solvent conditions, indicating a transition between these structures in the same region within a protein[47]. In the case of the p53 C-terminal IDD, a stretch of residues may form α-helix, ß-sheet or coil depending on the binding partner[48]. The work presented here builds on these phenomena and focuses on the unbound IDP partner, where local structural nuclei around leucine residues in E7N holds back structure formation which, depending on context or solvent, may go to either α-helix or ß-sheet type of structures.

Despite the conserved short interdomain distance in E7, both domains were so far not reported to interact in a structure or to form persistent contacts, and this may well be explained by the ID nature of E7N. The increased proline content in this stretch in combination with a slow isomerization we determined to take place in P47 as part of an epitope recognized by a monoclonal antibody[34] strongly suggest a functional role for this hinge region, as a sort of link of dynamic or transient interactions between the two domains. We have shown that most α-helix structure gain in response to mutation and TFE occurred within E7N, but E7N alone (E71-40) could not reach the same amount of α-helix in TFE[17,29]. This means that E7C is required for the incremented α-helix structure caused by mutation, and further supports the idea of at least transient domain interaction through long-range contacts, that could not be observed to date by any structural method. In addition, we have shown here that both domains are required for the formation of ß-structure.

In summary, hyperconserved residues within an intrinsically disordered domain stabilize local structures that oppose to α-helix folding or slow down ß-sheet aggregation, and mutation of these release structures by increasing α-helix content and accelerating aggregation. The fact that the conserved residues are distributed along the E7N IDD as opposed to contiguous in sequence, strongly suggest that these structures are local and nuclei-like. It is important to discriminate *aggregation* from *folding* nuclei: the aggregation nucleus refer to a partly folded oligomerization competent monomer or dimer[39], while the folding nucleus consists of local interactions of adjacent side chains within a polypeptide that are capable of guiding folding[39,49,50]. Figure 5 shows a proposed model scheme for an integrative analysis of our findings. E7 is a weak dimer, therefore in fast exchange with the monomer, and the E7N IDD does not participate of the dimerization interface, so we use the monomer as the starting point for a simplified representation. Anti-folding structural elements, possibly nuclei-like, formed around the conserved residues are shown as circles or knots, where the red one represents those affecting the ß-sheet route, and the green one those affecting α-helix folding. These do no exclude each other, something experimentally supported by the fact that L13A and P17A affect both routes (Table 1 and Table 2), and the mechanism by which they participate in each route is likely to differ, whether destabilizing antifolding nuclei, destabilizing the aggregation nucleus or stabilizing endpoint structures. The fact that the hyperconserved residues are mostly leucines and that they do not form α-helix nuclei, in fact they oppose to helix formation, suggest that the bulky aliphatic side chain is required for interactions at the local level.

The ß-sheet route can readily form a template which may lead to a discrete oligomer if the structure stabilized is interdomain and intramolecule, or an aggregate if the ß-sheet structure is stabilized intermolecular between dimers, generating a network that leads to a macroscopic precipitate (Fig. 5). In fact, we have shown that removal of zinc leads to soluble spherical oligomers rich in repetitive ß-sheet[20]. The α-helix route affecting nuclei is in fast exchange as is the case for helix folding in general, and most processes are reversible, unlike ß-sheet oligomerization and aggregation. Mutation releases α-helix, which may either form structures within the E7N IDD or fold against the globular E7C domain. Clearly, these are equilibria that are chemically possible and will determine the different ways of accommodating different binding partners, representing the essence of the concept of multitarget IDP recognition as an evolutive functional feature. They are very sensitive to mutation and it is hard to imagine that single mutations of residues involved in non-covalent interactions stabilizing the globular E7C will have such an impact on the overall fold. This could constitute an experimental demonstration that an IDE is at least as sensitive to mutation than a globular domain of the same protein subjected to the same evolutive pressure, and merits further investigation. The more sensitive to mutation, the highest pressure for particular residue conservation. This takes place at structural elements that oppose to folding, which unveil a yet undefined hidden structural code for intrinsic disorder only surfaced by mutagenesis guided in the dark by sequence conservation of a unique system. Our results clearly show that IPDs such as E7 are far from a random ensemble of conformers, and are a first step for establishing the structural basis of what may be lowly populated, dynamic or marginally stable structures, not accessible to direct physical measurement.

Funding Information. This work was funded by Instituto Nacional del Cáncer (2015–2017).

Acknowledgments: We thank C. Marino Busjle for comments on the manuscript. SSB, GC, LBC, MR, IES, LGA and GPG are career investigators from Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) and MGN was doctoral fellow from CONICET.

Supporting Information. Six additional figures: Sequence properties of HPV-16 E7 (Figure S1), Analysis of RNA structures using the RNAZ algorithm (Figure S2), 1H-15N HSQC spectrum of the HPV-16 E7 protein (Figure S3), Size exclusion chromatography (SEC) of E7wt and E7 IDD mutants (Figure S4), Comparative representation of the secondary structure of E7 IDD mutants (Figure S5), and pH dependence β-sheet formation and aggregation in E7 IDD mutants (Figure S6).

REFERENCES

1. van der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D. T., Kim, P. M., Kriwacki, R. W., Oldfield, C. J., Pappu, R. V., Tompa, P., Uversky, V. N., Wright, P. E., and Babu, M. M. (2014) Classification of intrinsically disordered regions and proteins, *Chem Rev 114*, 6589–6631.
2. Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W., Garner, E. C., and Obradovic, Z. (2001) Intrinsically disordered protein, *J Mol Graph Model 19*, 26–59.
3. Xue, B., Dunker, A. K., and Uversky, V. N. (2012) Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life, *J Biomol Struct Dyn 30*, 137–149.
4. Habchi, J., Tompa, P., Longhi, S., and Uversky, V. N. (2014) Introducing protein intrinsic disorder, *Chem Rev 114*, 6561–6588.
5. Wright, P. E., and Dyson, H. J. (2015) Intrinsically disordered proteins in cellular signalling and regulation, *Nat Rev Mol Cell Biol 16*, 18–29.
6. Babu, M. M. (2016) The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease, *Biochem Soc Trans 44*, 1185–1200.
7. Sormanni, P., Piovesan, D., Heller, G. T., Bonomi, M., Kukic, P., Camilloni, C., Fuxreiter, M., Dosztanyi, Z., Pappu, R. V., Babu, M. M., Longhi, S., Tompa, P., Dunker, A. K., Uversky, V. N., Tosatto, S. C., and Vendruscolo, M. (2017) Simultaneous quantification of protein order and disorder, *Nat Chem Biol 13*, 339–342.
8. Uversky, V. N., and Longhi, S. (2012) *Flexible viruses: Structural disorder in viral proteins*, John Wiley & Sons, Inc., Hoboken, New Jersey.
9. zur Hausen, H. (2009) Papillomaviruses in the causation of human cancers - a brief historical account, *Virology 384*, 260–265.
10. Moody, C. A., and Laimins, L. A. (2010) Human papillomavirus oncoproteins: pathways to transformation, *Nat Rev Cancer 10*, 550–560.
11. Chemes, L. B., Sanchez, I. E., Alonso, L. G., and de Prat-Gay, G. (2012) Intrinsic disorder in the human papillomavirus E7 protein, in *Flexible viruses: Structural disorder in viral proteins* (Uversky, V. N., and Longhi, S., Eds.), pp 313–346, John Wiley & Sons, Inc., Hoboken, New Jersey.
12. Munger, K., Phelps, W. C., Bubb, V., Howley, P. M., and Schlegel, R. (1989) The E6 and E7 genes of the human papillomavirus type 16 together are necessary and sufficient for transformation of primary human keratinocytes, *J Virol 63*, 4417–4421.
13. Alonso, L. G., Chemes, L. B., Cerutti, M. L., Dantur, K. I., and de Prat-Gay, G. (2012) Biochemical and structure-function analyses of the HPV E7 protein, in *Small DNA tumour viruses* (Gaston, K., Ed.), pp 99–124, Caister Academic Press, Norfolk, UK.
14. Chemes, L. B., Glavina, J., Alonso, L. G., Marino-Buslje, C., de Prat-Gay, G., and Sanchez, I. E. (2012) Sequence evolution of the intrinsically disordered and globular domains of a model viral oncoprotein, *PLoS One 7*, e47661.
15. Pim, D., and Banks, L. (2010) Interaction of viral oncoproteins with cellular target molecules: infection with high-risk vs low-risk human papillomaviruses, *APMIS 118*, 471–493.
16. Alonso, L. G., Garcia-Alai, M. M., Nadra, A. D., Lapena, A. N., Almeida, F. L., Gualfetti, P., and Prat-Gay, G. D. (2002) High-risk (HPV16) human papillomavirus E7 oncoprotein is highly stable and extended, with conformational transitions that could explain its multiple cellular binding partners, *Biochemistry 41*, 10510–10518.
17. Garcia-Alai, M. M., Alonso, L. G., and de Prat-Gay, G. (2007) The N-terminal module of HPV16 E7 is an intrinsically disordered domain that confers conformational and recognition plasticity to the oncoprotein, *Biochemistry 46*, 10405–10412.
18. Liu, X., Clements, A., Zhao, K., and Marmorstein, R. (2006) Structure of the human Papillomavirus E7 oncoprotein and its mechanism for inactivation of the retinoblastoma tumor suppressor, *J Biol Chem 281*, 578–586.
19. Ohlenschlager, O., Seiboth, T., Zengerling, H., Briese, L., Marchanka, A., Ramachandran, R., Baum, M., Korbas, M., Meyer-Klaucke, W., Durst, M., and Gorlach, M. (2006) Solution structure of the partially folded high-risk human papilloma virus 45 oncoprotein E7, *Oncogene 25*, 5953–5959.
20. Alonso, L. G., Garcia-Alai, M. M., Smal, C., Centeno, J. M., Iacono, R., Castano, E., Gualfetti, P., and de Prat-Gay, G. (2004) The HPV16 E7 viral oncoprotein self-assembles into defined spherical oligomers, *Biochemistry 43*, 3310–3317.

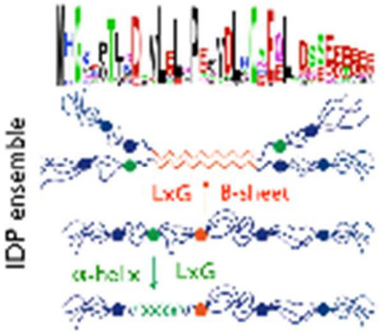
21. Alonso, L. G., Smal, C., Garcia-Alai, M. M., Chemes, L., Salame, M., and de Prat-Gay, G. (2006) Chaperone holdase activity of human papillomavirus E7 oncoprotein, *Biochemistry 45*, 657–667.
22. Dantur, K., Alonso, L., Castano, E., Morelli, L., Centeno-Crowley, J. M., Vighi, S., and de Prat-Gay, G. (2009) Cytosolic accumulation of HPV16 E7 oligomers supports different transformation routes for the prototypic viral oncoprotein: the amyloid-cancer connection, *Int J Cancer 125*, 1902–1911.
23. Smal, C., Alonso, L. G., Wetzler, D. E., Heer, A., and de Prat Gay, G. (2012) Ordered self-assembly mechanism of a spherical oncoprotein oligomer triggered by zinc removal and stabilized by an intrinsically disordered domain, *PLoS One 7*, e36457.
24. Chemes, L. B., Glavina, J., Faivovich, J., de Prat-Gay, G., and Sanchez, I. E. (2012) Evolution of linear motifs within the papillomavirus E7 oncoprotein, *J Mol Biol 422*, 336–346.
25. Parigi, G., Rezaei-Ghaleh, N., Giachetti, A., Becker, S., Fernandez, C., Blackledge, M., Griesinger, C., Zweckstetter, M., and Luchinat, C. (2014) Long-range correlated dynamics in intrinsically disordered proteins, *J Am Chem Soc 136*, 16201–16209.
26. Ozenne, V., Schneider, R., Yao, M., Huang, J. R., Salmon, L., Zweckstetter, M., Jensen, M. R., and Blackledge, M. (2012) Mapping the potential energy landscape of intrinsically disordered proteins at amino acid resolution, *J Am Chem Soc 134*, 15138–15148.
27. Abascal, F., Zardoya, R., and Telford, M. J. (2010) TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations, *Nucleic Acids Res 38*, W7–W13.
28. Gruber, A. R., Neubock, R., Hofacker, I. L., and Washietl, S. (2007) The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures, *Nucleic Acids Res 35*, W335–W338.
29. Noval, M. G., Gallo, M., Perrone, S., Salvay, A. G., Chemes, L. B., and de Prat-Gay, G. (2013) Conformational dissection of a viral intrinsically disordered domain involved in cellular transformation, *PLoS One 8*, e72760.
30. Chemes, L. B., Camporeale, G., Sanchez, I. E., de Prat-Gay, G., and 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*, 1680–1696.
31. Chemes, L. B., Alonso, L. G., Noval, M. G., and de Prat Gay, G. (2012) Circular dicroism techniques for the analysis of intrinsicaly disordered proteins and domains, in *Intrinsically disordered protein analysis* (Uversky, V. N., and Dunker, A. K., Eds.), pp 387–404, Springer protocols.
32. Jasanoff, A., and Fersht, A. R. (1994) Quantitative determination of helical propensities from trifluoroethanol titration curves, *Biochemistry 33*, 2129–2135.
33. Monsellier, E., Ramazzotti, M., de Laureto, P. P., Tartaglia, G. G., Taddei, N., Fontana, A., Vendruscolo, M., and Chiti, F. (2007) The distribution of residues in a polypeptide sequence is a determinant of aggregation optimized by evolution, *Biophys J 93*, 4382–4391.
34. Fassolari, M., Chemes, L. B., Gallo, M., Smal, C., Sanchez, I. E., and de Prat-Gay, G. (2013) Minute time scale prolyl isomerization governs antibody recognition of an intrinsically disordered immunodominant epitope, *J Biol Chem 288*, 13110–13123.
35. Li, X., Johansson, C., Cardoso Palacios, C., Mossberg, A., Dhanjal, S., Bergvall, M., and Schwartz, S. (2013) Eight nucleotide substitutions inhibit splicing to HPV-16 3'-splice site SA3358 and reduce the efficiency by which HPV-16 increases the life span of primary human keratinocytes, *PLoS One 8*, e72776.
36. Grassmann, K., Rapp, B., Maschek, H., Petry, K. U., and Iftner, T. (1996) Identification of a differentiation-inducible promoter in the E7 open reading frame of human papillomavirus type 16 (HPV-16) in raft cultures of a new cell line containing high copy numbers of episomal HPV-16 DNA, *J Virol 70*, 2339–2349.
37. Pace, C. N., and Scholtz, J. M. (1998) A helix propensity scale based on experimental studies of peptides and proteins, *Biophys J 75*, 422–427.
38. Fersht, A. R. (1998) Kinetics in protein folding, in *Structure and mechanism in protein science: A guide to enzyme catalysis and protein folding* (Fersht, A. R., Ed.), pp 540–570, Freeman.
39. Frieden, C. (2007) Protein aggregation processes: In search of the mechanism, *Protein Sci 16*, 2334–2344.
40. Anfinsen, C. B. (1973) Principles that govern the folding of protein chains, *Science 181*, 223–230.
41. Borysik, A. J., Kovacs, D., Guharoy, M., and Tompa, P. (2015) Ensemble Methods Enable a New Definition for the Solution to Gas-Phase Transfer of Intrinsically Disordered Proteins, *J Am Chem Soc 137*, 13807–13817.
42. Tompa, P. (2012) Intrinsically disordered proteins: a 10-year recap, *Trends Biochem Sci 37*, 509–516.
43. Fersht, A. R. (1998) Protein Engineering, in *Structure and mechanism in protein science: A guide to enzyme catalysis and protein folding* (Fersht, A. R., Ed.), pp 420–449, Freeman.
44. Miller, S., Janin, J., Lesk, A. M., and Chothia, C. (1987) Interior and surface of monomeric proteins, *J Mol Biol 196*, 641–656.
45. Makhatadze, G. I., and Privalov, P. L. (1995) Energetics of protein structure, *Adv Protein Chem 47*, 307–425.
46. Minor, D. L., Jr., and Kim, P. S. (1996) Context-dependent secondary structure formation of a designed protein sequence, *Nature 380*, 730–734.

47. Cordes, M. H., Burton, R. E., Walsh, N. P., McKnight, C. J., and Sauer, R. T. (2000) An evolutionary bridge to a new protein fold, *Nat Struct Biol 7*, 1129–1132.
48. Oldfield, C. J., Meng, J., Yang, J. Y., Yang, M. Q., Uversky, V. N., and Dunker, A. K. (2008) Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners, *BMC Genomics 9 Suppl 1*, S1.
49. Daggett, V., and Fersht, A. R. (2003) Is there a unifying mechanism for protein folding?, *Trends Biochem Sci 28*, 18–25.
50. Fersht, A. R. (1998) Folding pathways and energy landscapes, in *Structure and mechanism in protein science: A guide to enzyme catalysis and protein folding* (Fersht, A. R., Ed.), pp 573–611, Freeman.

FOR TABLE OF CONTENTS USE ONLY

Hidden Structural Codes in Protein Intrinsic Disorder

Silvia S. Borkosky, Gabriela Camporeale, Lucía B. Chemes, Marikena Risso, María Gabriela Noval,
Ignacio E. Sánchez, Leonardo G. Alonso, and Gonzalo de Prat Gay*


IDP ensemble
LxG
β-sheet
α-helix
LxG