# A tetramerization domain in prokaryotic and eukaryotic transcription regulators homologous to p53

Nerea Bernardo,[a]‡ Isidro Crespo,[a]‡ Anna Cuppari,[a] Wilfried J. J. Meijer[b] and D. Roeland Boer[a]*

[a]Experiments Division, ALBA Synchrotron Light Source, Carrer de la Llum 2–26, 08290 Cerdanyola del Vallès, Catalunya, Spain, and [b]Centro de Biología Molecular 'Severo Ochoa' (CSIC–UAM), Universidad Autónoma de Madrid, Calle Nicolás Cabrera 1, Canto Blanco, 28049 Madrid, Spain. *Correspondence e-mail: rboer@cells.es

Transcriptional regulation usually requires the action of several proteins that either repress or activate a promotor of an open reading frame. These proteins can counteract each other, thus allowing tight regulation of the transcription of the corresponding genes, where tight repression is often linked to DNA looping or cross-linking. Here, the tetramerization domain of the bacterial gene repressor Rco from *Bacillus subtilis* plasmid pLS20 ($Rco_{pLS20}$) has been identified and its structure is shown to share high similarity to the tetramerization domain of the well known p53 family of human tumor suppressors, despite lacking clear sequence homology. In $Rco_{pLS20}$, this tetramerization domain is responsible for inducing DNA looping, a process that involves multiple tetramers. In accordance, it is shown that $Rco_{pLS20}$ can form octamers. This domain was named $TetD_{loop}$ and its occurrence was identified in other *Bacillus* species. The $TetD_{loop}$ fold was also found in the structure of a transcriptional repressor from *Salmonella* phage SPC32H. It is proposed that the $TetD_{loop}$ fold has evolved through divergent evolution and that the $TetD_{loop}$ originates from a common ancestor predating the occurrence of multicellular life.

## 1. Introduction

Proper gene regulation is essential for every organism to adjust the expression profile of its encoded genes to changing environmental conditions. The mechanisms of transcriptional regulation are very different in prokaryotes and eukaryotes, and are generally more complex in the latter. For example, in eukaryotes the genomes are packed in more sophisticated ways and transcriptional regulators rely less on sequence specificity for DNA response elements (REs; Youssef *et al.*, 2019). The reason for this higher level of sophistication is most likely related with intercellular communication in multi-cellular organisms, which becomes clear when considering situations in which gene regulation is disturbed and causes disease (Lee & Young, 2013). The well studied human tumor suppressor protein p53 illustrates the importance of proper transcriptional regulation (Vogelstein *et al.*, 2000; Lane & Levine, 2010). Mutations in this protein have been associated with the occurrence of tumorigenesis for over four decades (Rivlin *et al.*, 2011; Perri *et al.*, 2016).

The active control of the conformation of the DNA duplex regulates access to promoter regions and thereby makes an important contribution to gene regulation. An example of this
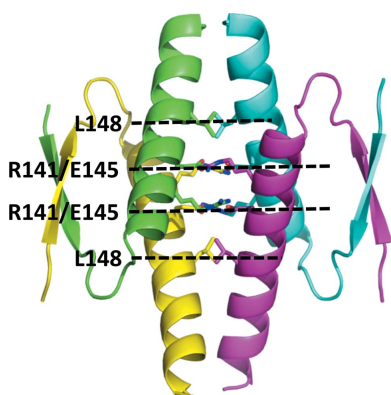
is long-range dsDNA looping, which is not to be confused with the formation of short hairpin loops in ssDNA. In long-range looping, two DNA elements are brought into close proximity by introducing a kink or strong bend in the DNA located between the two elements. DNA looping occurs frequently in both prokaryotes as well as eukaryotes (Cournac & Plumbridge, 2013; Vilar & Saiz, 2005; Morelli et al., 2009). Examples of transcriptional regulation through DNA looping in prokaryotes are found in the metabolic genes *ara*, *gal*, *lac* and *deo* and in phage systems (for a review, see Matthews, 1992). In addition, the human tumor suppressor protein p53, which represents a large family of homologs in metazoa, induces the DNA bending required for binding (McKinney & Prives, 2002). Interestingly, all of these regulators form homotetramers and have been shown to be essential for DNA looping, even though the folds of the proteins involved do not share any structural similarity. This suggests that the tetrameric quaternary structure provides a particular functional advantage in the formation of DNA loops, which may be related to the cooperativity in the stability conferred by having two DNA-recognition anchoring points at both extremes of the loop region.

p53 has been studied extensively and many functions have been attributed to this protein in processes related to, amongst others, human development and DNA repair and metabolism (Gaglia et al., 2013; Lane & Levine, 2010). It consists of five domains: two transcription-activation domains (TADs; residues 1–40 and 40–61), a proline-rich region (residues 64–92), a DNA-binding domain (DBD; residues 98–303), a nuclear localization signal-containing region (residues 303–323), an oligomerization domain (residues 323–365) and a C-terminal basic domain (residues 363– 393). The TADs are important for the transactivation of different target genes. The DBD recognizes the p53 recognition element (p53RE), which consists of two copies of a 5′-RRRCWWGYYY-3′ sequence (IUPAC nomenclature, where R is A or G, Y is C or T and W is A or T) separated by a spacer of 0–13 bp (Brázda & Coufal, 2017; El-Deiry et al., 1992). The oligomerization domain induces tetramerization and possibly other oligomerization states of p53, which has been amply documented (reviewed in Chène, 2001). We will refer to this domain as p53Tet. The C-terminal domain is highly charged due to a high lysine content and is intrinsically disordered. This domain has a dual function: the positive charge enhances the affinity for DNA and at the same time is responsible for recruiting other factors. p53 is also able to form DNA loops and link DNA across large distances due to the combined action of the DBD and the p53Tet domain, which itself does not bind DNA but rather links DBD-bound DNAs (Stenger et al., 1994; Kearns et al., 2016; Brázda & Coufal, 2017). Many of the oncogenic mutations occur in the DBD (Perri et al., 2016), but several have been mapped to the tetramerization domain (reviewed in Chène, 2001; see references in Kamada et al., 2011; Petitjean et al., 2007). In addition, mutations in p53Tet cause Li–Fraumeni syndrome, a hereditary disease that conveys a high disposition to develop early-onset neoplasms (Etzold et al., 2015).

Here, we report a structural analog of the oligomerization domain of p53 from the conjugative plasmid pLS20 of the Gram-positive (G+) bacterium *Bacillus subtilis*. This transcriptional regulator, named Rco (Singh et al., 2013), forms tetramers in solution (Ramachandran et al., 2014; Crespo et al., 2020; Singh et al., 2013) and represses the main conjugation promoter $P_c$ that controls the transcription of genes 28–74. The promotor contains 11 copies of the sequence 5′-CAGTGAAA-3′ and variations thereof (Ramachandran et al., 2014), which are likely to be the recognition sites for $Rco_{pLS20}$. $Rco_{pLS20}$ controls its own expression by regulating the activity of the overlapping and divergently oriented $P_r$ promoter (reviewed in Meijer et al., 2021). The simultaneous regulation of the $P_c$ and $P_r$ promoters involves $Rco_{pLS20}$-mediated DNA looping, which is achieved through the binding of $Rco_{pLS20}$ to two operators separated by 75 bp and located near the $P_c$ and $P_r$ promoters, each containing multiple direct repeats of the Rco-recognition element (rcoRE; Ramachandran et al., 2014). One of the operators, $O_{II}$, overlaps with the $P_c$ and $P_r$ promoters and contains at least six direct repeats of the rcoRE. The other operator, $O_I$, contains at least four direct repeats of the rcoRE, which are convergently oriented with respect to the rcoREs in operator $O_{II}$. DNA recognition by Rco is most likely to be achieved through a conserved helix–turn–helix (HTH) domain at its N-terminus, which is followed by a sequence of hitherto unknown function at its C-terminus. Given the presence of at least ten rcoREs in the promotor region to which Rco binds, it can be expected that at least two Rco tetramers bind to this region.

We have previously solved the structure of $Rap_{pLS20}$ (Crespo et al., 2020), which is the response regulator that binds $Rco_{pLS20}$, thereby activating the conjugation promotor and allowing expression of the conjugation operon (Singh et al., 2013). The aim of this study was to understand the structural mechanism of transcriptional regulation of $Rco_{pLS20}$. For this purpose, we further analyzed the oligomerization behavior of $Rco_{pLS20}$ and identified the oligomerization domain. We confirm the formation of tetramers and show that $Rco_{pLS20}$ can also form octamers under specific conditions. Furthermore, we present the crystal structure of the tetramerization domain of $Rco_{pLS20}$. The domain encompasses 35 residues of the C-terminal region and is structurally homologous to the tetramerization domain of the human oncogene p53. As this fold is implicated in the formation of DNA loops, we designate domains that have this fold $TetD_{loop}$. We define a motif for $TetD_{loop}$ and suggest that the $TetD_{loop}$ domain is ubiquitous among all kingdoms of life.

The implications of the occurrence of $TetD_{loop}$ in prokaryotes and eukaryotes are profound. First of all, the structures suggest that the fold precedes the appearance of multicellular life, which had not been considered (see, for example, Joerger et al., 2014). This implies that these proteins share a common ancestor and hence are ubiquitous. Secondly, it suggests that a basic paradigm of gene regulation through DNA looping exists among prokaryotes and eukaryotes, and we believe that the possible parallels between the mechanism of action of p53 and $Rco_{pLS20}$ should be further investigated.

## 2. Materials and methods

### 2.1. Production and purification of Rco$_{pLS20}$

Cloning, expression, isolation and purification of Rco$_{pLS20}$ were performed as described previously (Crespo *et al.*, 2020). Typically, a yield of 20 mg Rco$_{pLS20}$ was obtained from 10 g cell pellet. Purity was assessed to be >95% by SDS–PAGE followed by Coomassie Blue staining. Protein concentration was determined from the absorbance at 280 nm on a Nanodrop 2000 spectrophotometer (ThermoFisher Scientific) using an extinction coefficient $\varepsilon$(1%) of 2.93. Protein was used for assays immediately where possible or stored in aliquots at −80°C.

### 2.2. Crystallization of RcoTet$_{pLS20}$

Rco$_{pLS20}$ was concentrated to 17 mg ml$^{-1}$ using an Amicon Ultra 15 ml centrifugal filter (Merck Millipore) with a cutoff of 10 kDa in a buffer consisting of 250 m$M$ NaCl, 20 m$M$ Tris pH 8.0. Concentrated Rco$_{pLS20}$ was gently mixed with a previously annealed double-stranded oligonucleotide (Biomers.net, Germany) with forward sequence 5′-GTCAGTGAAAAA-3′ in a 2:1 (protein:DNA) stoichiometry. The crystals giving the highest resolution data were obtained by the sitting-drop vapor-diffusion method at 18°C by equilibration of drops consisting of 100 nl protein solution and 100 nl crystallization buffer [0.1 $M$ HEPES pH 7.5, 28%($v/v$) PEG 600] against 100 µl crystallization buffer in the reservoir. The crystals took three months to grow and were harvested for X-ray diffraction data collection by cryocooling them by direct transfer from the crystallization drop into liquid nitrogen.

### 2.3. Data collection and processing

Data collection was performed on the BL13-XALOC beamline at the ALBA Synchrotron Light Source at 100 K. Data were processed with *AutoPROC* (Vonrhein *et al.*, 2011) using anisotropic resolution cutoffs (see Table 1). The structure was determined *de novo* using *ARCIMBOLDO* (Rodríguez *et al.*, 2009) followed by automated model building in *Phenix* version 1.12-2829 (Adams *et al.*, 2010). The structure was refined using *Phenix* interspersed with manual adjustments in *Coot* version 0.8.9.2205 (Emsley *et al.*, 2010). The refinement statistics are given in Table 1. The structure was deposited in the PDB with accession code 8bny.

### 2.4. Calculation of an *AlphaFold* model of full-length Rco$_{pLS20}$

*AlphaFold* version 2.1.0 (Jumper *et al.*, 2021) was used to generate five models of full-length Rco$_{pLS20}$ (UniProt entry E9RIY8). The five models are essentially the same and the highest ranking model was used for analysis.

### 2.5. Database searches and figures

*BlastP* (Altschul *et al.*, 1990) searches were performed using residues Ser124–Asp161 of the Rco$_{pLS20}$ sequence. Figures were prepared using *PyMOL* version 2.3 (Schrödinger). Superpositions were performed using the built-in 'align' function in *PyMOL*. *PDBeFold* from EMBL–EBI was used to identify folds similar to that of Rco$_{pLS20}$ in the PDB (Krissinel & Henrick, 2005).

### 2.6. Size-exclusion chromatography (SEC) assays

25 µg Rco$_{pLS20}$ (25 µl) were injected into a Superdex 200 Increase 5/150 GL column (GE Healthcare) that had been equilibrated with buffers at different pH values. For pH 5, the column was equilibrated with 500 m$M$ NaCl, 20 m$M$ citrate buffer pH 5. For pH 8, 500 m$M$ NaCl, 20 m$M$ Tris pH 8 was used. For pH 10, 500 m$M$ NaCl, 20 m$M$ glycine–NaOH pH 10 was used. Elution was performed at a flow rate of 0.2 ml min$^{-1}$. The elution was continuously monitored at wavelengths of 280 and 260 nm. Estimation of the molecular weight (MW) was performed from the elution volume $V_{el}$ of the detected peaks using an in-house calibration of the relation between log(MW) and $V_{el}$ of proteins with known MW. The equation derived from this calibration was $V_{el} = -0.6815\log(MW) + 5.1906$, with $R^2 = 0.933$.

## 3. Results

### 3.1. Crystal structure of the tetramerization domain of Rco$_{pLS20}$

Crystallization trials on apo full-length (FL) Rco$_{pLS20}$ were not successful, but cocrystallization of FL Rco$_{pLS20}$ with

**Table 1**
Summary of the data-processing and refinement statistics for crystallographic analysis of the RcoTet$_{pLS20}$ structure (PDB entry 8bny).

Values in parentheses are for the outer resolution shell.

| Data collection | |
|---|---|
| Beamline | XALOC, ALBA |
| Wavelength (Å) | 0.9793 |
| Space group | $P2_12_12_1$ |
| $a$, $b$, $c$ (Å) | 35.33, 36.18, 109.53 |
| Resolution range (Å) | 34.36–1.409 (1.433–1.409) |
| No. of reflections | |
| Total | 96338 (5832) |
| Unique | 18842 (943) |
| Ellipsoidal completeness (%) | 91.6 (51.9) |
| $\langle I/\sigma(I)\rangle$ | 14.5 (1.7) |
| Average multiplicity | 5.1 (6.2) |
| $R_{merge}$† (%) | 3.9 (71.4) |
| $R_{meas}$‡ (%) | 4.4 (78.0) |
| CC$_{1/2}$ (%) | 99.9 (51.2) |
| Structure refinement | |
| $R_{cryst}$§/$R_{free}$¶ (%) | 23.32/24.91 |
| R.m.s.d. from target values | |
| Bond lengths (Å) | 0.0058 |
| Bond angle distances (Å) | 0.724 |
| *MolProbity* scores | |
| Clashscore (‰) | 2.0 |
| Poor rotamers (%) | 2.10 |
| Ramachandran outliers (%) | 0.00 |
| Ramachandran favored (%) | 97.0 |
| Overall score (%) | 1.41 |
| Isotropic $B$-factor analysis | |
| Average model $B$ factors (Å$^2$) | 38.49 |
| $B$ factor from Wilson plot (Å$^2$) | 35.0 |
| No. of non-H atoms | 1263 |
| No. of solvent molecules | 62 |

† $R_{merge} = \sum_{hkl}\sum_i |I_i(hkl) - \langle I(hkl)\rangle|/\sum_{hkl}\sum_i I_i(hkl)$. ‡ $R_{meas} = \sum_{hkl}\{N(hkl)/[N(hkl) - 1]\}^{1/2}\sum_i |I_i(hkl) - \langle I(hkl)\rangle|/\sum_{hkl}\sum_i I_i(hkl)$. § $R_{cryst} = \sum_{hkl}||F_{obs}| - |F_{calc}||/\sum_{hkl}|F_{obs}|$. ¶ $R_{free} = \sum_{hkl\in T}||F_{obs}| - |F_{calc}||/\sum_{hkl\in T}|F_{obs}|$, where $T$ represents a test set comprising ~5% of all reflections that were excluded during refinement.

various DNA sequences did result in the formation of crystals. However, structure determination revealed that the crystals contained only part of Rco$_{pLS20}$, corresponding to residues Val125–Lys159 of the FL protein, and no DNA molecules (Fig. 1). Since the missing sequence of the protein cannot fit into the asymmetric unit, it is likely that the protein degraded during the course of the crystallization experiment, which often occurs in multidomain proteins. Indeed, size-exclusion chromatography (SEC) analysis of the protein confirmed that degradation occurred over time (Supplementary Fig. S1), leading to the appearance of fragments with a molecular weight compatible with the crystallized fragment. We show below that the crystallized fragment indeed corresponds to one of the domains of Rco$_{pLS20}$, which we will refer to as RcoTet$_{pLS20}$. The asymmetric unit contains four monomers forming one crystallographically independent tetramer.

Each monomer consists of an elongated sequence comprising residues Val125–Thr132, which includes a four-amino-acid $\beta$-strand formed by residues Arg127–Asp130. This strand is followed by a sharp turn facilitated by the glycine residue Gly133, which is followed by an $\alpha$-helix comprising residues Glu136–Lys157 (Fig. 1a). The RcoTet$_{pLS20}$ tetramer consists of a dimer of primary dimers. The primary homodimer is formed by the arrangement of two $\beta$-strands from two monomers in an antiparallel fashion and by concomitant antiparallel packing of the helices against one face of the two $\beta$-strands (Fig. 1c). The $\alpha$-helices interact with one of the faces of the $\beta$-strands through hydrophobic interactions involving Val131, Phe129, Leu134, Ile139, Val142, Ile146 and Leu149 (Fig. 1b). At the center of the exposed helical face, a cluster of charged residues is formed by Arg141 and Glu145. The hydrophobic Leu148 residues are located at both extremes of this cluster along the $\alpha$-helix (Figs. 1c and 1d), which connects to the hydrophobic cluster through Leu149. The tetramer is formed through interactions of the Arg141, Glu145 and Leu148 residues, which we will refer to as the REL motif, from the four monomers (Figs. 1c and 1d). The carboxylic acid groups of the Glu145 residues interact through hydrogen bonds (Fig. 1d).

The overall shape of the tetrameric structure is reminiscent of an octagon (see Fig. 1e, left panel). The C-terminal ends of two helices form two pairs of opposed vertices of the octagon. The remaining four vertices are formed by the N-termini of the $\beta$-strands. Thus, two pairs of N-terminal DBDs are located on opposing sides of the RcoTet$_{pLS20}$ domain. The lateral edges of the octagon formed by the $\beta$-strands are hydrophobic (Fig. 1e), which is expected to contribute to interactions with the DBDs based on analogy with other structures (see below). The planar faces of the octagonal box are formed by $\alpha$-helices (Fig. 1e).

### 3.2. pH-dependent oligomerization behavior of Rco$_{pLS20}$

Rco$_{pLS20}$ tetramerization seems to mainly be driven by the charged interactions of the REL motif (Figs. 1c and 1d). This triggered us to study the oligomerization behavior of full-length Rco$_{pLS20}$ at pH 5, pH 8 and pH 10, respectively. We found that Rco$_{pLS20}$ tends to form higher order oligomers under alkaline conditions (pH 10; Fig. 1f). Under acidic conditions, i.e. pH 5, disruption of the tetramer is observed (Fig. 1f). It is likely that protonation-induced neutralization of the carboxylates of the central Glu145 occurs at pH 5. This causes disruption of the counter-charge stabilized hydrogen-bonding network, shown in Fig. 1(d), between the agglomerated Glu145 and Arg141 residues at the tetramerization interface.

The MW of the species that form at pH 5, pH 8 and pH 10 were estimated by calibration of the SEC column elution based on the individual elution of a set of proteins of distinct MW under equivalent conditions (Fig. 1f). The MW estimates are 56.99 kDa at pH 5, 85.43 kDa at pH 8 and 151.57 kDa at pH 10 (Supplementary Table S1). Given that the MW of FL Rco$_{pLS20}$ is 20.32 kDa, the elution peaks therefore correspond to two to three FL protein molecules at pH 5, four protein molecules at pH 8 and eight protein molecules at pH 10 (Supplementary Table S1). It is unlikely that Rco$_{pLS20}$ can form a trimer at pH 5; it is far more likely that protonation of the Glu145 residues at pH 5 disrupts the tetramerization interface, resulting in dimers. The slight deviation in the elution pattern from a dimer at pH 5 may be a result of an altered surface charge under acidic conditions, which may affect interactions with the agarose–dextran matrix of the chromatography column, or a mixture of dimers and tetramers that results in an average migration of the peak. These data show that the protein prefers forming octamers at pH 10 even at low concentrations, which can be isolated by SEC. No direct evidence for hexamers, heptamers or complexes larger than octamers have been observed by SEC under the conditions tested.

### 3.3. Comparison of the TetD$_{loop}$ folds of structural homologs

A search of the PDB (Berman et al., 2000) for structural homologs of RcoTet$_{pLS20}$ using PDBeFold (Krissinel & Henrick, 2005) resulted in several significant hits, which all corresponded to the human oncogene p53 and its analogs (Supplementary Table S2). The basic fold of all these structures consists of a pair of short $\beta$-strands followed by an $\alpha$-helix connected by a kinked loop (Figs. 1a and 2 and Supplementary Fig. S2). Hydrophobic residues line the internal surfaces of the $\beta$-strands and the $\alpha$-helix, thereby forming the hydrophobic core of the structure. Both the $\beta$-strands and the $\alpha$-helix have similar lengths, and structural differences are mainly found in the angle between the $\beta$-strands and the $\alpha$-helix, which appears to be conditioned by the residues forming the kinked loop. In the structures identified and analyzed here this angle is smaller than 27° when a glycine is present in the loop, whereas it ranges from 34° to 62° when this glycine is lacking, as is the case for Drosophila melanogaster p53 (Dmp53) and CEP-1 (Supplementary Fig. S2). Due to the variation in the angle, the monomers of the structural homologs generally do not superpose, which complicates the detection of structural similarity. Strikingly, the structure of Rep$_{SPC32H}$ (Kim et al., 2016), a repressor

MGNREQFDLSKY LLEKRKQKGLSQTQVANDTGLSSAYISMLEKGERKRPTQAVIKKLATALSIPNIEMI QIMEYIASMEE
--- HHHHHH | HHHHHHHHH | --- HHHHHHH --- HHHHH | HH ------- | HHHHHHHH | ---- HHHHH | --------- |
　　　　10　　　　　　20　　　　　　　　　　40　　　　　　50　　　　　　60　　　　　　70　　　　　　80

DHVKEKDKDTSKQIKLFDLTGLTEASINQIQNQIDQLKSSENNS VIRFFDVTGLSEKDIERVKEEIELLKIRNEYMKLK KD
-------- HHHHHHHH --- | --- HHHHHHH | HHHHHHHHHHHHHHHHH | ----- HHHH | HHHHHHHH | HHHHHHH -- | -
　　　　90　　　　100　　　　110　　　　120　　　　130　　　　140　　　　150　　　　160
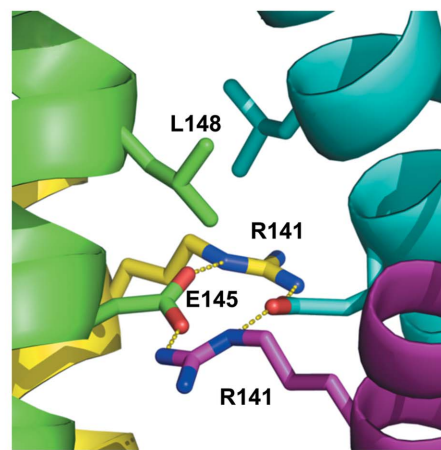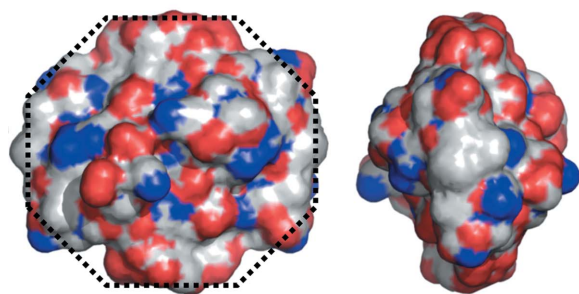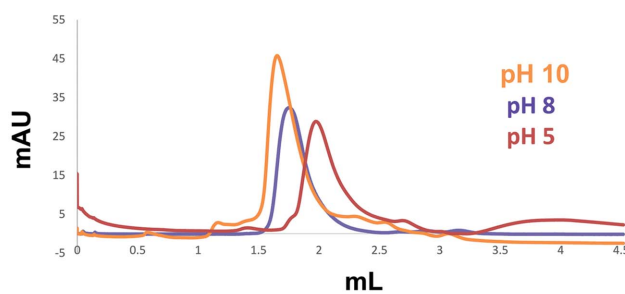
*(a)*
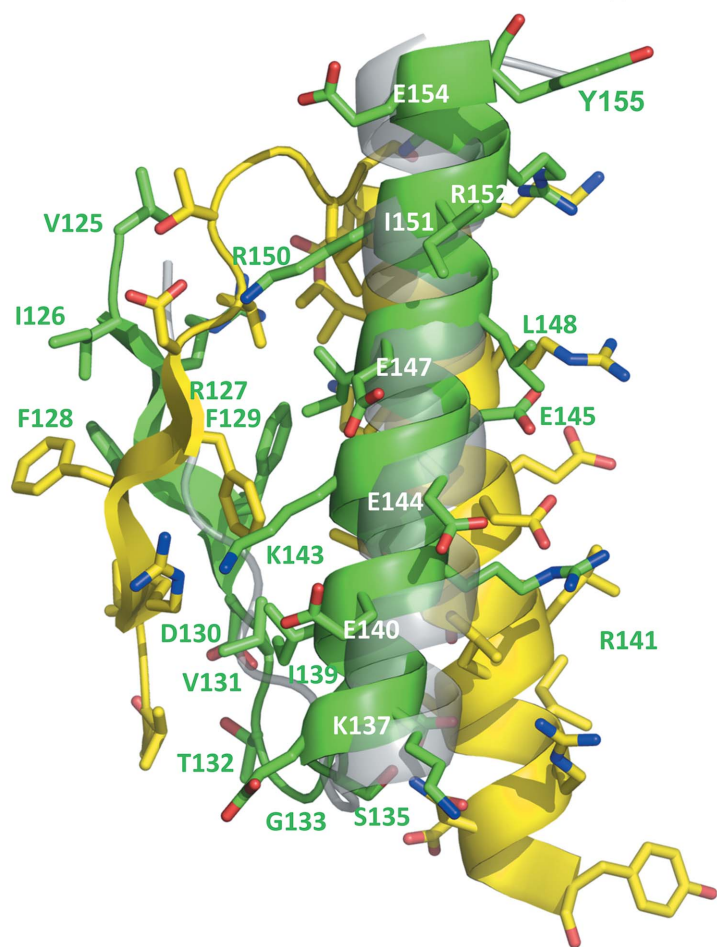


*(b)*
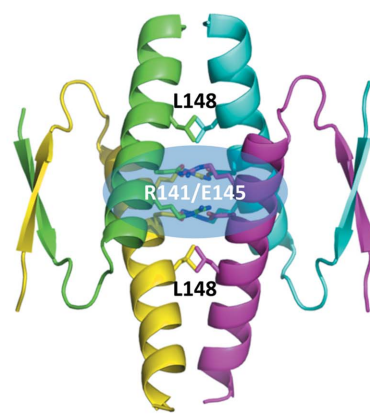


*(c)*



*(d)*



*(e)*



*(f)*

**Figure 1**
Overview of the structure and oligomerization behavior of RcoTet$_{pLS20}$. (*a*) The sequence and secondary-structure prediction using *JPred*4 (Drozdetskiy *et al.*, 2015) of full-length Rco$_{pLS20}$. The rectangles indicate the HTH domain at the N-terminus as annotated in UniProt entry E9RIY8 and the RcoTet$_{pLS20}$ domain corresponding to the crystallized fragment. The amino-acid letters are colored according to secondary structure in the crystal structure (green for β-strand and red for α-helix). (*b*) Cartoon representation of residues Val125–Tyr155 of the RcoTet$_{pLS20}$ dimer, colored by chain. All side chains are shown as sticks; the numbers of selected residues of the green monomer are indicated. A monomer of the p53 structure (PDB entry 1aie; Mittl *et al.*, 1998), superposed on the green monomer of Rco$_{pLS20}$, is shown in transparent gray. (*c*) Side view of the tetrameric structure, showing the two charged layers formed by the Arg141 and Glu145 residues from the four monomers and the way each of these clusters is capped by a pair of Leu148 residues. (*d*) Close-up view of a single Arg141/Glu145 layer and Leu148 cap. The hydrogen bonds between the residues of the charged cluster are indicated by yellow dotted lines. (*e*) Electrostatic potential (ESP) on the surface of the tetrameric structure of RcoTet$_{pLS20}$ in two orientations. The orientation shown in the left panel is similar to that shown in Fig. 1(*b*). In the right panel, RcoTet$_{pLS20}$ is rotated by 90° around a vertical axis, showing the ESP of the view on the exposed face of the β-strands. (*f*) Superposed SEC elution profiles of RcoTet$_{pLS20}$ at different pH values.

encoded by *Salmonella* phage SPC32H, reveals a tetramerization domain that is structurally similar to RcoTet$_{pLS20}$ and p53Tet. This domain is also called CAD and, apart from inducing tetramers in Rep$_{SPC32H}$, it is also responsible for its interaction with the antirepressor Ant (Kim *et al.*, 2016). We propose that it is a TetD$_{loop}$ and will refer to it as such here. This structure did not appear in the *PDBeFold* search using RcoTet$_{pLS20}$ as a query as described above. Instead, Rep$_{SPC32H}$ was identified as an Rco$_{pLS20}$ analog based on the following shared features. Firstly, the architecture of the two full-length proteins is similar and contains a DBD at the N-terminus and a TetD$_{loop}$ at the C-terminus. Rep$_{SPC32H}$ additionally contains a dimerization domain (MDD) situated between the DBD and the TetD$_{loop}$ (Kim *et al.*, 2016). Secondly, the DBDs are homologous in sequence. Thirdly, Rep$_{SPC32H}$ is a transcriptional repressor and, together with its antirepressor protein Ant, exhibits a similar regulatory mechanism to the Rap$_{pLS20}$–Rco$_{pLS20}$ pair. Thus, Rep$_{SPC32H}$ prevents entry into the lytic cycle of the phage by tight repression of the genes essential for the lytic cell cycle. However, the structural mechanism of gene repression of this protein has not been extensively characterized.

### 3.4. Relative spatial position of the TetD$_{loop}$ and additional domains

The angle between the $\beta$-strands and the $\alpha$-helix in the monomer ultimately determines the respective orientations of the dimers in the quaternary structure of the tetramer. This further complicates the detection of structural similarity between structural homologs. The interface of the tetramer is all hydrophobic for p53 (Fig. 2$b$) and all of its homologs, except for Dmp53, which has a charged core similar to RcoTet$_{pLS20}$, consisting of interacting glutamate residues lined with arginines in a two-layered configuration, as shown for RcoTet$_{pLS20}$ in Figs. 1 and 2($a$). Interestingly, the $\beta$-strands and the

$\alpha$-helix are duplicated in the sequence of Dmp53, and the tetramer interface is therefore formed by eight helices, *i.e.* two helices from each of the four monomers. The exact nature of the amino acids involved in the hydrophobic tetramerization interfaces can vary, which is exemplified by Rep$_{SPC32H}$ and p53: the former uses phenylalanines Phe187 and Phe190 of the four monomers to form this interface, whereas in p53 a cluster of leucine and methionine residues is found.

The quaternary structures of the TetD$_{loop}$ of p53 and Rco$_{pLS20}$ are similar (Fig. 2), despite the difference in the relative orientations of the dimers across the tetramerization interface. The configuration of the interaction between primary dimers causes the N-termini of the TetD$_{loop}$ to point in opposite directions. This is confirmed by the FL structure of a complex of p53 and p300 determined by electron microscopy (Ghosh *et al.*, 2019), which shows a planar overall structure with the DBDs interacting pairwise; each pair extends away from the central p53Tet domain in opposite directions (Fig. 3$a$). The placement of the MDDs in the structure of Rep$_{SPC32H}$ is similar (Fig. 3$b$), where two pairs of adjacent MDDs are separated by $\sim$100 Å, as measured between the far ends of these domains. The highest ranked *AlphaFold*2 model of a tetramer of full-length Rco (Supplementary Fig. S3$a$) is consistent with the arrangement of the RHH domains of p53 and Rep$_{SPC32H}$ shown in Figs. 3($a$) and 3($b$), respectively. It is interesting to note that the *AlphaFold*2 model of the tetramer of the full-length protein (Supplementary Fig. S3$b$) shows that the tetramerization domain is well predicted for this oligomerization state.

### 3.5. Structurally similar proteins show low sequential homology

The vertebrate p53 homologs generally share low sequence homology in regions other than their DNA-binding domains (Ou *et al.*, 2007; Lu & Abrams, 2006). In line with this,
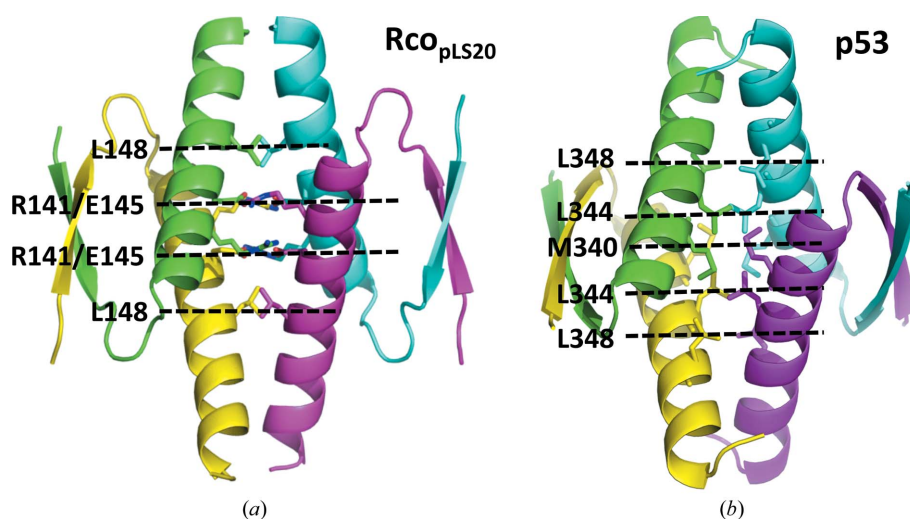


**Figure 2**
Structure of tetramer-forming domains of Rco$_{pLS20}$ and human p53. ($a$) Cartoon representation of the structure of the tetramerization domain of Rco$_{pLS20}$ (this work; PDB entry 8bny). ($b$) Cartoon representation of the structure of the tetramerization domain of p53 (PDB entry 1aie). The hydrophobic residues forming the interface are indicated.

RcoTet$_{pLS20}$ and Rep$_{SPC32H}$ also share low homology with these proteins. For example, the residues conferring interactions between the $\beta$-strands and the $\alpha$-helix in the core of the dimer are hydrophobic, but the identities of these residues are not conserved between the different TetD$_{loop}$ domains. Similarly, even within vertebrates the tetramerization interface has diverged among different proteins, leading to the occurrence of hydrophobic and/or charged tetramerization interfaces as described above. It is therefore not surprising that sequential homology is low in these structurally similar proteins.

Despite low sequence conservation, we used the RcoTet$_{pLS20}$ domain (Rco$_{pLS20}$ residues Ser124–Asp161) as a query in *BlastP* searches. As expected from the low sequence homology, none of the structural homologs described above were identified in the *BlastP* search. Instead, nine nonredundant proteins with *E*-values ranging between $3 \times 10^{-6}$ and $2 \times 10^{-3}$ were identified, which are all encoded by bacteria belonging to the phylum *Firmicutes* (Fig. 3c). Thus, the TetD$_{loop}$ is found in G+ proteins related to *B. subtilis*.

## 4. Discussion

Rco$_{pLS20}$ is a transcriptional regulator that exerts its function through the formation of a DNA loop by binding to two regions that are separated by about 75 bp (Ramachandran *et al.*, 2014). The N-terminal HTH motif identified in the N-terminal region is likely to be involved in recognition of the rcoREs, but the structural basis of looping was not well understood up to this point. Our present results show that the master regulator Rco$_{pLS20}$, which is crucial in the transcriptional regulation of the conjugation operons of plasmid pLS20, contains a tetramerization domain that is likely to be involved in formation of the DNA loop, and that this domain has high structural homology to the tetramerization domain present in the p53 family of tumor repressors. The structure reveals that RcoTet$_{pLS20}$ forms tetramers like the p53 tumor suppressor protein family and shows how the different DNA-recognition elements are bridged. Identification of the tetramerization domain based on sequence was hampered by its low sequence homology with similar domains of known function.
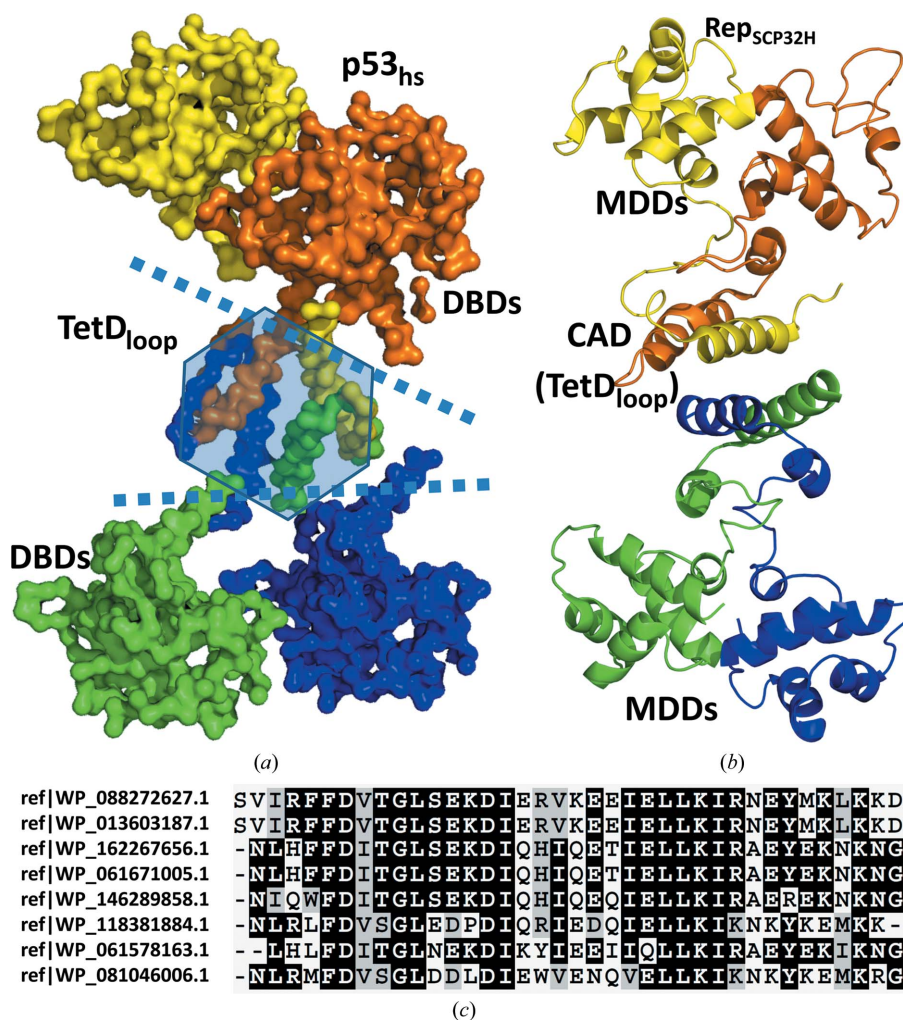


**Figure 3**
TetD$_{loop}$ structures identified to be analogs of RcoTet$_{pLS20}$. (a) $C^\alpha$ trace of the tetrameric electron microscopy structure (PDB entry 5xzc; Ghosh *et al.*, 2019) of full-length human p53 complexed with p300 (not included). The four monomers are colored differently. The DBD and tetramerization domains are labeled and the boundaries between these domains are indicated by dotted lines. (b) Cartoon representation of one of the tetramers from the structure of the phage repressor Rep$_{SPC32H}$ (PDB entry 5d4z). (c) Multiple sequence alignment of the *B. subtilis* sequences retrieved by a *BlastP* search of the RcoTet$_{pLS20}$ domain.

Rco$_{pLS20}$ is one of many examples illustrating the intimate link between protein tetramerization and DNA looping or long-range DNA cross-linking observed in both prokaryotic and eukaryotic systems. In prokaryotes, regulators have been extensively studied, for example *gal* and *lac* (Cournac & Plumbridge, 2013; Matthews, 1992), whereas in eukaryotes the p53 analogs (Kearns *et al.*, 2016; Stenger *et al.*, 1994) are well known examples. To date, however, no structural homology has been reported between the structures of tetrameric, DNA-loop inducing proteins from eukaryotes and non-eukaryotes. Remarkably, we found that the TetD$_{loop}$ domain observed in Rco$_{pLS20}$ and p53 is also present in the structure of *Salmonella* phage SPC32H (PDB entry 5d4z). However, structural homology of this domain to existing structures was not detected in this study (Kim *et al.*, 2016), although a structural comparison with PDB structures using *DALI* was performed. It is likely that variations in the sequence and in the relative orientation of the $\beta$-strands and $\alpha$-helix hampered detection. In fact, the Rep$_{SPC32H}$ linker consists of a short $\alpha$-helix which is not present in any of the other structures. This connecting $\alpha$-helix allows the $\beta$-strands and the second $\alpha$-helix to adopt a nearly parallel configuration, without the need for a glycine residue. Furthermore, there is no sequence homology between the dimer and tetramer interface of the TetD$_{loop}$ of Rep$_{SPC32H}$, and all the proteins analyzed above.

Despite the low sequence homology, it was possible to discern motifs for the proteins that contain a charged tetramerization interface. In our analysis, a positively charged residue (Arg141 in RcoTet$_{pLS20}$) should be followed approximately a full turn later in the $\alpha$-helix (*i.e.* three or four positions downstream) by a glutamate (Glu145 in RcoTet$_{pLS20}$). In addition, we find a hydrophobic residue at a second full turn (+7) from the positively charged residue (Leu148 in RcoTet$_{pLS20}$). We named this distinctive motif the 'REL motif'. Furthermore, a glycine residue is often found in the sharp turn between the $\alpha$-helix and $\beta$-strands and may be of importance, as it is highly conserved in metazoan homologs of human p53 and in hits found through sequence-homology searches. From our comparison, it seems that a glycine is a prerequisite for a small angle between the $\beta$-strands and $\alpha$-helix, as a lack of this residue leads to differences in the orientation of the helices and $\beta$-strands of the respective monomers. It should be noted that variations on this motif exist, as the Rep$_{SPC32H}$ structure contains a short additional $\alpha$-helix perpendicular to the preceding $\beta$-strand and the trailing $\alpha$-helix (Fig. 3*b*). Variability is also observed for the tetramerization interface joining two dimers, which has evolved to either a completely hydrophobic interface or an interface stabilized by complementary charges.

Rco$_{pLS20}$ has the capacity to form higher order complexes as shown by the results described here and by SAXS analysis of FL Rco$_{pLS20}$ at neutral pH, where it was shown to occur in a concentration-dependent manner (Crespo *et al.*, 2020). The pH dependence of octamerization determined here suggests that this interface is also charged, like that of the tetramer. Analysis of the distribution of charges and hydrophobic patches across the surface of RcoTet$_{pLS20}$ suggests that the interaction is likely to be mediated by the charged face of the octagonal form of the tetramer, given that the edges are hydrophobic. Remarkably, inspection of the interactions between symmetry-related molecules of the crystal structure of Rep$_{SPC32H}$ reveals interactions between the TetD$_{loop}$ of adjacent tetramers, resulting in octamers that are in accordance with the model of octamerization through the TetD$_{loop}$, as proposed for Rco$_{pLS20}$. The model that emerges from the DNA binding of the tetramer is that of cooperative binding of recognition sites across the DNA loop. The initial RcoTet$_{pLS20}$–DNA complex would be formed stochastically, perhaps with the aid of helper proteins that shape the DNA at the turn and through the intrinsic propensity of the DNA to bend at the loop region (Ramachandran *et al.*, 2014). Binding of several tetramers of the repressor across the loop stabilizes the loop structure.

The striking structural similarity between the prokaryotic and eukaryotic tetramerization domains exemplified by RcoTet$_{pLS20}$ and p53, respectively, raises the question whether they evolved independently or whether they are derived from a common ancestor. This question cannot be answered straightforwardly, as sequence homology in these short sequences is difficult to detect. However, the conservation of the REL motif in some interfaces, coupled with the structural similarity of the core strand–kink–helix, suggests common ancestry. For example, RcoTet$_{pLS20}$ and Dmp53 are structurally homologous and both have a charged tetramerization interface, albeit that duplication of the secondary-structure elements occurs in Dmp53 as described above. In addition, a well established cofactor of p53, named Strap, consists of a TPR motif (Adams *et al.*, 2012) and interacts with p53 (Jung *et al.*, 2007). Thus, Strap could be a functional homolog of Rap$_{pL20}$.

The combination of a shared core structure and tetramerization interface and cofactors of similar structure suggests that convergent evolution is unlikely and strongly favors the hypothesis that all proteins incorporating a TetD$_{loop}$ have diverged from a common ancestor. The presence of the TetD$_{loop}$ motif across the gene pools of all kingdoms of life also supports divergent evolution, as it is unlikely to have occurred through multiple convergent events. We propose therefore that all members of this family stem from a common ancestor that included DBD and TetD$_{loop}$ domains. This would imply that the TetD$_{loop}$ fold predates the occurrence of multicellularity some 3–3.5 billion years ago (Grosberg & Strathmann, 2007).

The work described here suggests a common regulatory mechanism that exists in all kingdoms of life and that controls gene transcription, which is mediated by a tetramerization domain that we call TetD$_{loop}$. The detection of this domain in sequences and structures is complicated due to low sequence conservation and to structural variation of the constituting elements. We describe the motifs that we have observed in the different structures. However, improved bioinformatics tools that can increase the predictive power for identifying small domains such as TetD$_{loop}$ would help in the detection of these

domains in proteins that share similar mechanisms conferred by this domain.

## References

Adams, C. J., Pike, A. C. W., Maniam, S., Sharpe, T. D., Coutts, A. S., Knapp, S., La Thangue, N. B. & Bullock, A. N. (2012). *Proc. Natl Acad. Sci. USA*, **109**, 3778–3783.

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). *Acta Cryst.* D**66**, 213–221.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* **215**, 403–410.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Brázda, V. & Coufal, J. (2017). *Int. J. Mol. Sci.* **19**, 2737.

Chène, P. (2001). *Oncogene*, **20**, 2611–2617.

Cournac, A. & Plumbridge, J. (2013). *J. Bacteriol.* **195**, 1109–1119.

Crespo, I., Bernardo, N., Miguel-Arribas, A., Singh, P. K., Luque-Ortega, J. R., Alfonso, C., Malfois, M., Meijer, W. J. J. & Boer, D. R. (2020). *Nucleic Acids Res.* **48**, 8113–8127.

Drozdetskiy, A., Cole, C., Procter, J. & Barton, G. J. (2015). *Nucleic Acids Res.* **43**, W389–W394.

El-Deiry, W. S., Kern, S. E., Pietenpol, J. A., Kinzler, K. W. & Vogelstein, B. (1992). *Nat. Genet.* **1**, 45–49.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.

Etzold, A., Schröder, J. C., Bartsch, O., Zechner, U. & Galetzka, D. (2015). *Fam. Cancer*, **14**, 161–165.

Gaglia, G., Guan, Y., Shah, J. V. & Lahav, G. (2013). *Proc. Natl Acad. Sci. USA*, **110**, 15497–15501.

Ghosh, R., Kaypee, S., Shasmal, M., Kundu, T. K., Roy, S. & Sengupta, J. (2019). *Biochemistry*, **58**, 3434–3443.

Grosberg, R. & Strathmann, R. (2007). *Annu. Rev. Ecol. Evol. Syst.* **38**, 621–654.

Joerger, A. C., Wilcken, R. & Andreeva, A. (2014). *Structure*, **22**, 1301–1310.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. (2021). *Nature*, **596**, 583–589.

Jung, H., Seong, H.-A. & Ha, H. (2007). *J. Biol. Chem.* **282**, 35293–35307.

Kamada, R., Nomura, T., Anderson, C. W. & Sakaguchi, K. (2011). *J. Biol. Chem.* **286**, 252–258.

Kearns, S., Lurz, R., Orlova, E. V. & Okorokov, A. L. (2016). *Nucleic Acids Res.* **44**, 6185–6199.

Kim, M., Kim, H. J., Son, S. H., Yoon, H. J., Lim, Y., Lee, J. W., Seok, Y.-J., Jin, K. S., Yu, Y. G., Kim, S. K., Ryu, S. & Lee, H. H. (2016). *Proc. Natl Acad. Sci. USA*, **113**, E2480–E2488.

Krissinel, E. & Henrick, K. (2005). *Computational Life Sciences*, edited by M. R. Berthold, R. C. Glen, K. Diederichs, O. Kohlbacher & I. Fischer, pp. 67–78. Berlin, Heidelberg: Springer.

Lane, D. & Levine, A. (2010). *Cold Spring Harb. Perspect. Biol.* **2**, a000893.

Lee, T. I. & Young, R. A. (2013). *Cell*, **152**, 1237–1251.

Lu, W.-J. & Abrams, J. M. (2006). *Cell Death Differ.* **13**, 909–912.

Matthews, K. S. (1992). *Microbiol. Rev.* **56**, 123–136.

McKinney, K. & Prives, C. (2002). *Mol. Cell. Biol.* **22**, 6797–6808.

Meijer, W. J. J., Boer, D. R., Ares, S., Alfonso, C., Rojo, F., Luque-Ortega, J. R. & Wu, L. J. (2021). *Front. Mol. Biosci.* **8**, 648468.

Mittl, P. R. E., Chène, P. & Grütter, M. G. (1998). *Acta Cryst.* D**54**, 86–89.

Morelli, M. J., ten Wolde, P. R. & Allen, R. J. (2009). *Proc. Natl Acad. Sci. USA*, **106**, 8101–8106.

Ou, H. D., Löhr, F., Vogel, V., Mäntele, W. & Dötsch, V. (2007). *EMBO J.* **26**, 3463–3473.

Perri, F., Pisconti, S. & Della Vittoria Scarpati, G. (2016). *Ann. Transl. Med.* **4**, 522.

Petitjean, A., Mathe, E., Kato, S., Ishioka, C., Tavtigian, S. V., Hainaut, P. & Olivier, M. (2007). *Hum. Mutat.* **28**, 622–629.

Ramachandran, G., Singh, P. K., Luque-Ortega, J. R., Yuste, L., Alfonso, C., Rojo, F., Wu, L. J. & Meijer, W. J. J. (2014). *PLoS Genet.* **10**, e1004733.

Rivlin, N., Brosh, R., Oren, M. & Rotter, V. (2011). *Genes Cancer*, **2**, 466–474.

Rodríguez, D. D., Grosse, C., Himmel, S., González, C., de Ilarduya, I. M., Becker, S., Sheldrick, G. M. & Usón, I. (2009). *Nat. Methods*, **6**, 651–653.

Singh, P. K., Ramachandran, G., Ramos-Ruiz, R., Peiró-Pastor, R., Abia, D., Wu, L. J. & Meijer, W. J. J. (2013). *PLoS Genet.* **9**, e1003892.

Stenger, J. E., Tegtmeyer, P., Mayr, G. A., Reed, M., Wang, Y., Wang, P., Hough, P. V. & Mastrangelo, I. A. (1994). *EMBO J.* **13**, 6011–6020.

Vilar, J. M. & Saiz, L. (2005). *Curr. Opin. Genet. Dev.* **15**, 136–144.

Vogelstein, B., Lane, D. & Levine, A. J. (2000). *Nature*, **408**, 307–310.

Vonrhein, C., Flensburg, C., Keller, P., Sharff, A., Smart, O., Paciorek, W., Womack, T. & Bricogne, G. (2011). *Acta Cryst.* D**67**, 293–302.

Youssef, N., Budd, A. & Bielawski, J. P. (2019). *Methods Mol. Biol.* **1910**, 3–31.