THE EFFECTS OF PRIMARY SEQUENCE PERTURBATION ON THE STRUCTURE OF INTRINSICALLY DISORDERED PROTEINS

by

Benjamin J. Ricard, B.S.

A thesis/dissertation submitted to the Graduate Council of
Texas State University in partial fulfillment
of the requirements for the degree of
Master of Science
with a Major in Biochemistry
December 2017

Committee Members:

Steve Whitten, Chair

Karen Lewis

Sean Kerwin

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplication Permission

DEDICATION

To my parents, Melissa and Richard Ricard, to whom I express extreme gratitude for their patience and support of both my scientific career and my personal life.

ACKNOWLEDGEMENTS

I would like to extend my gratitude to Dr. Steve Whitten for providing me with training in scientific rigor and allowing me to freely explore my scientific ideas. The experience I've gained in his lab will prove doubtlessly invaluable in my future endeavors as a scientist, and as a human. I am also thankful to my committee members Dr. Karen Lewis and Dr. Sean Kerwin for their guidance and advice. I am thankful to all of my fellow lab members, particularly Lance English and Leona Martin, for their patience in teaching me techniques necessary for this project. Taylor Perrin deserves my thanks, particularly for assisting in purifying RL11p53. Finally, I am deeply grateful for the support from my friends and family without which nothing, not even the beauty of the natural world and the mathematics contained within, would be worth exploring.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF ILLUSTRATIONS

Illustration **Page**

LIST OF ABBREVIATIONS

Abbreviation	Description
2xYT	16 g/L Tryptone, 10 g/L Yeast Extract, 5.0g/L NaCl
A	Alanine (amino acid sequences) or Adenine (DNA sequences)
Ampicillin-r	Ampicillin resistance gene, produces β-lactamase
BME	2-mercaptoethanol
C	Cysteine (amino acid sequences) or Cytosine (DNA sequences)
C11p53	p53(1-93) with 11 prolines closest to C-terminus substituted to glycine
CD	Circular Dichroism
D	Aspartate
DEAE	DEAE-Sepharose
df	Degrees of freedom
DNA	Deoxyribonucleic Acid
DNP-Aspartate	N-2,4-DNP-L-aspartic acid
DOS	Density of States
E	Glutamate
E_r	Energy of general microstate *r*
F	F-statistic
F	Phenylalanine
FASTA	FAST-All

G.....Glycine (amino acid sequences) or Guanine (DNA sequences)

GLY .....Glycine

GuHCl.....Guanidinium hydrochloride

GWAS.....Genome-wide association studies

H.....Histidine

HDM2 .....E3 ubiquitin-protein ligase Mdm2

HIS .....Histidine, or denoting the sequence of six histidines used for thrombin cleavage

I .....Isoleucine

IDP .....Intrinsically Disordered Protein

IDR.....Intrinsically Disordered Region

IPTG.....Isopropyl β -D-1-thiogalactopyranoside

K.....Lysine

K_a .....Number of non-synonymous mutations

k_b .....Boltzmann's Constant, $1.38064 \cdot 10^{-23}$ m^2 kg s^{-2} K^{-1}

K_d .....Retention factor for SEC

kDa..... Kilodalton

K_s.....Number of synonymous mutations

L .....Leucine

lacI .....Gene for lacI, the lac repressor protein

lacO.....Binding site for lac repressor protein

lacO1.....Binding site for lac repressor protein

LB ..... Lysogeny broth

M ..... Methionine

M11p53 ..... p53(1-93) with 11 middle-most prolines substituted to glycine

MRE ..... Molar residual ellipticity

ANOVA ..... Analysis of variants

mRNA ..... Messenger RNA

MS ..... Mean of squares

N ..... Asparagine

N11p53 ..... p53(1-93) with 11 prolines closest to N-terminus substituted to glycine

NaAc ..... Sodium Acetate

NMR ..... Nuclear Magnetic Resonance

Ori_pUC ..... Bacterial origin of replication

P ..... p-value

P ..... Proline

P_Amp ..... Promoter for the Ampicillin gene

P_lacI ..... Promoter of lacI gene

P_T5_Inducible ..... Promoter sequence for the T5 bacteriophage RNA polymerase

P_T7_Inducible ..... Promoter sequence for the T7 bacteriophage RNA polymerase

p53 ..... Tumor suppressor protein p53

p53(1-93) ..... First 93 residues of p53

p53(93-1) ..... p53(1-93) transcribed from C-terminus to N-terminus

PDB ..... Protein Data Bank

PP_{II} ..... Polyproline-II helix

p_r ..... Probability of microstate *r*

PRO ..... Proline

PrP^c ..... Human prion protein

PSI ..... Pounds per square inch

Q ..... Glutamine

R ..... Arginine

RBS ..... Ribosome binding sequence

RF11p53 ..... N11p53 transcribed from C-terminus to N-terminus

R_h ..... Hydrodynamic Radius

RL11p53 ..... C11p53 transcribed from C-terminus to N-terminus

RM11p53 ..... M11p53 transcribed from C-terminus to N-terminus

RNA ..... Ribonucleic Acid

RPM ..... Rotations per minute

S ..... Serine

SDS ..... sodium dodecyl sulfate

SDS ..... sodium dodecyl sulfate polyacrylamide gel electrophoresis

SEC ..... Size Exclusion Chromatography

SOC ..... 2% tryptone, 0.5% yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM MgSO4, and 20 mM glucose

SS ..... Sum of squares

T ..... Threonine (amino acid sequences) or Thymine (DNA sequences)

TAD ..... Transactivation Domain

Term_bla ..... transcription termination sequence for *Ampicillin-r*

Term_rpoC ..... Transcription termination sequence for rpoC

Term_T7 ..... Termination sequence for bacteriophage T7 polymerase

TGS ..... 25 mM TrisHCl, .192 M Glycine, 0.1% SDS

TP53 ..... Gene encoding p53

Tris-HCl ..... Tris(hydroxymethyl)aminomethane hydrochloride

V ..... Valine

V_0 ..... Elution time of Blue Dextran

V_e ..... Elution time of sample

V_t ..... Elution time of DNP-Aspartate

W ..... Tryptophan

Y ..... Tyrosine

Z ..... Partition Function

β ..... $1/k_bT$

ΔA ..... Differential Absorption

ABSTRACT

Intrinsically disordered proteins (IDPs) are a class of proteins that do not converge to a set of similar, energetically stable tertiary folds, but rapidly fluctuate between a wide variety of different accessible energy states. Previous work established the dependence of hydrodynamic radius (R_h) on intrinsic conformational propensity, particularly that of the polyproline-II (PP_{II}) helix, which is in turn dependent on the primary sequence. To explore this relationship between primary sequence and structure, six IDPs having the same number and type of each amino acid derived from the intrinsically disordered N-terminus of p53, were created by substituting 11 prolines of the total 22 prolines, which has a high propensity to form the PP_{II} helix, to glycine, which has a low propensity to form the PP_{II} helix, in different locations; the N-terminus, the C-terminus, and the middle of the protein, as well as their retro-transposed variants, i.e. translating from the original C-terminus to the original N-terminus. Additionally, the wild-type p53, as well as its retro-transposed variant, was isolated and purified to further test direct directional dependence of structure. The propensity to form the PP_{II} helix is measured by circular dichroism spectroscopy, and R_h is measured through size exclusion chromatography. Results showed a significant ($p < 0.05$) dependence of structure on directionality, as measured by differences in R_h; furthermore, circular dichroism spectra suggests that this collapse in structure is due to reduced occupancy of the PP_{II} helix. Both directionality and location of amino acids appear to modulate the structure of IDPs, indicating local charge-driven effects.

I. INTRODUCTION

Overview

Understanding the structure and function of biological macromolecules is of importance in elucidating the life-sustaining mechanisms of the natural world and the molecular origins leading to diseased states. Proteins, a polymer made of amino acids linked together through peptide bonds, perform a variety of functions in biological organisms. Understanding the physical three-dimensional fold of proteins provides extensive information regarding their energetic landscape, which is a map of physical conformation to thermodynamic properties [1]. Such fundamental understanding of the generalized thermodynamic laws of proteins can be used in the creation of novel protein structures for a variety of industrial purposes, as well as open avenues for potential protein-targeted drug design. The traditional paradigm for understanding proteins, holds that the primary sequence dictates the secondary structure, and the secondary structure determines the tertiary structure [2]. However, this model fails to provide meaningful context for understanding the structure of intrinsically disordered proteins (IDPs), which are proteins that do not have stable tertiary folds. As a result, it is challenging to analyze structures that are not suited for crystallization via x-ray diffraction, which is the gold standard in structural determination of biological macromolecules [3].

Intrinsically disordered proteins are defined as proteins that rapidly fluctuate between a wide variety different *microstates*, or energetically accessible conformations, in contrast to *folded proteins*, which converge towards a system of more closely related microstates. Around 33% of all eukaryotic proteins are estimated to be disordered or contain intrinsically

disordered regions (IDRs) [4]. IDPs play diverse roles in cells, particularly in nucleic acid and protein recognition. Genome-wide association studies (GWAS) have shown that protein mutations involved in human cancers are statistically enriched with IDPs [5]. Prediction of intrinsic disorder can be made based solely on properties observed to be enriched in IDPs, such as the number of charged and hydrophilic amino acids, regardless of sequence position, as IDPs lack the extensive, site-specific tertiary contacts relative to folded proteins [6, 7]. Though the traditional structure-function paradigm of proteins holds that tertiary structure dictates the functionality of proteins, these disordered proteins do not contain a fixed tertiary structure and still retain functionality [8]. From a physical perspective, statistical-thermodynamic stability still exists within IDPs in the cellular environment [9]. Despite lack of a rigid structure, IDPs can adjust to a state of lower energy due to the system of dynamic equilibrium of distinct conformations. IDPs can adjust to a state of lower energy in a given environment than a protein more constrained to only accessing a smaller number of energetic states. [10]. An example of a common functional motif of IDPs that a folded protein cannot mimic is a coupled binding-folding mechanism, in which binding of a ligand by an IDP induces a conformation change stabilizing the IDP into a folded state [8]. This transition is analogous to, but fundamentally different than, binary structural changes such as the T to R transition in hemoglobin as a result of oxygen binding [11]. Instead of converting from one stable conformation to another, however, the IDP moves between a wide variety of ensembles. Many fundamental processes in nature depend heavily on the structure of IDPs and other amorphous polymers, and these flexible segments are uniquely able to compliment the certain roles of biological processes that structures with tertiary stability cannot [12].

From an experimental standpoint, direct understanding of the properties of both the general IDP and specific ones (in this study, the N-terminus of p53) provide a new avenue to target these biologically relevant polymers for treatment of disease, targeted drug design, antibody design, and other practical applications [13-15]. Furthermore, theoretical understanding of these IDPs pushes the boundary of current knowledge in many different fields, such as biophysics, structural biology, and evolutionary biology, as due to fundamentally different properties than traditionally studied folded proteins, IDPs open a new aperture for tying together interdisciplinary problems. For example, the evolution trajectories of IDPs have different properties than evolution of folded proteins, such as decreased sequence coevolution, increased positive selection pressures measured through K_a/K_s ratio, the ratio of nonsynonymous (K_a) to synonymous (K_s) mutation rates, and functionary robustness to highly divergent sequence evolution in similarly functional proteins [16-18]. More directly, examination of physical properties of IDPs can further insight to the general rules that govern the structure of universal proteins. Current analysis of proteins, largely taking place without consideration to disordered structures is inherently incomplete, and a unification of the frameworks that govern both disordered and ordered protein structures is currently an underexplored area of molecular biology.

Many IDPs are involved in ligand recognition and binding [19]. Understanding the discrete influences of *global* (properties of protein chains that hold under permutation independent combinations of amino acids) and *local* (properties of protein chains that are dependent on the permutation of amino acids) regulation on IDP structure will provide understanding on the mechanisms that IDPs use to bind potential targets. Intriguing properties of DNA binding proteins, many of which are disordered, include sequence non-

specificity but DNA ligand global structural specificity, which highlights the importance of understanding the role that the global properties of architectural proteins play in DNA topology modulation [20]. Potential applications include a novel avenue for drug design in targeting specific ligands by differentially modulating global and local sequence effects.

Structural Properties of IDPs and the Polyproline-II helix

The PP_{II} is an extended left-handed helix with dihedral angles φ = -75° and ψ = +145°, commonly observed in fractional IDP ensembles, and is observed in IDPs to exhibit noncooperativity, due to the lack of native intramolecular tertiary contacts [21]. The polyproline-II helix conformation had been initially recognized in the structure of the collagen triple-helix [22]. Collagen represents one-third of the total protein in humans, and takes the fibrous structure of three parallel polyproline-II helices coiled into a right-handed helix [23]. The three individual strands of collagen repeat the motif X-Y-Gly, where X and Y can represent any amino acid, but is dominated by either proline (28%) or 4-hydroxyproline (38%). Later, it was discovered that the conformation could exist in non-fibrous proteins that did not consist of a statistical overrepresentation of proline, such as beta-melanocyte-stimulating hormone [24]. Due to the expanded polyproline-II helix conformation, proteins with higher PP_{II} propensities tend to have larger hydrodynamic radius, or tumbling volume in water, than proteins with more canonical structures such as α-helix and β-sheet, explaining why IDPs tend to be larger than others, with similar number of residues[25]. Previous work has shown structure perturbations in the hydrodynamic radius (R_h) can be predicted from

changes in PP_{II} propensities, and these predictions show good agreement to experimentally measured values [25].

More recently, it has been shown that IDPs have significant intrinsic structural bias for the PP_{II} helix, indicating that it is a significant structural element of many IDPs[22, 26, 27]. Evolutionary conservation of PP_{II} bias has been observed in IDPs[28]. Previous work has established the extensive dependence of hydrodynamic radius (R_h) on sequence-dependent conformational propensity, particularly for the PP_{II} and α-helices, based on the PP_{II} propensity scale proposed by Hilser et. al [25, 29]. The implications that IDPs can be predicted, with good agreement, based only on intrinsic PP_{II} propensity indicate the statistical bias in the structure of IDPs, which are significantly larger than what would be predicted if IDPs took a random coil conformation, with relatively small correction for charge effects by measurement of charge groups laying within the Debye length of the given solvent [25].

The R_h provides a simple yet physically meaningful metric to understand the structural compaction, inductively, the degree of solvent-chain interactions of IDPs and thus intrinsic conformational propensities, structural properties, and physicochemical features of IDPs. Prior work indicates the R_h of IDPs is also responsive to changes in other global variables, particularly net charge density and temperature[25, 30]. Previous work established the following model from data obtained from 22 intrinsically disordered proteins, predicting the R_h of IDPs based on the number of residues, propensity for the polyproline-II helix from Elam et. al and a linear net-charge correction term, given by[25]:

$$R_h = 2.16\, N^{0.503-0.11ln\,(1-f_{ppii})} + 0.25\, Q - 0.31 N^{0.5}$$

The model disregards sequence position, and instead considers PP_{II} as an averaged sum of each fractional PP_{II}. The model indicates that the intrinsic conformational propensities are dependent on global factors alone. We directly test this hypothesis by creating six IDPs with identical inputs into the model for R_h.

Protein Polymer Physics and Fractals

Many of the properties, such as noncooperativity of PP_{II} and the power-law model of R_h of IDPs make them amenable for characterization through *fractals*. A *fractal* is an infinite iteration of self-similar portions, existing at every scale. A *polymer* is a chain of connected *monomer* building blocks; herein for example, amino acids are monomer subunits of a polymer protein chain. For a polymer, the scaling of hydrodynamic radius and peptide length follow a power-law scaling relationship of the form [31]

$$R_h \sim N^v$$

Where R_h represents the hydrodynamic radius, N represents the number of peptides, and v represents a term sometimes referred to as the "Flory exponent" that is dependent on solvation conditions [32, 33]. An *ideal polymer* is a polymer that has equal chain-chain and chain-solvent interaction energies, also known as a *Θ solvent*. An *excluded volume polymer*, or a polymer in a "*good" solvent,* is a polymer-solvent system that favors higher chain-solvent over chain-chain interactions. Conversely, a "*poor" solvent* is a solvent that has

favorable chain-chain interactions, rather than chain-solvent interactions. An ideal solvent has a Flory exponent $\nu = ½$, and Flory exponents are increased in "good" solvents, and decreased in "poor" solvents. Previous work has shown that IDPs can be modeled as excluded volume polymers with Flory exponents $\nu > ½$ [25]. Physically, this quantity is represented by the observed larger hydrodynamic sizes for equivalent values of N in IDPs and folded proteins. In "poor" solvent-polymer systems, intrachain interactions dominate over solvent-chain effects. Relevant examples of "poor" polymer-solvent interactions include that of the collapsed or globule state of the general folded protein [32].

Self-similarity is property of geometric objects where the larger scale structure of that object is similar to a smaller scale portion of the object. One relevant example of this behavior is a *fractal.* The property of a parameter to be similar at different scales is known as *scale-invariance*. The degree of a structure to follow a fractal pattern is given by the fractal dimension D, and is given by the reciprocal of the Flory exponent:

$$D = \frac{1}{\nu}$$

The fractal dimension D represents the (possibly noninteger) degree to which a polymer changes as the scale changes [32]. Thus, an excluded volume polymer-solvent system has a higher D than a "poor" solvent system. A decreasing fractal dimension implies less changing at different scales, i.e. the global structure is more similar to a smaller segment of the structure, while an increasing fractal dimension implies less similarity to a smaller segment [32]. Alternatively, a smaller fractal dimension implies that a polymer can more readily be

explained in terms of global structure, as opposed to local structure, as the details of local interactions are offset in comparison to global effects.

We can represent the accessible energetic states of proteins through a classical, discrete partition function of the form [34];

$$Z = \sum_{r} e^{-\beta E_r}$$

Where β represents the reciprocal of the multiple of Boltzmann's constant k_b and temperature in Kelvin, and E_r represents the energy iterating over each accessible microstate *r*. The probability p_r of obtaining a state in a heat bath at constant temperature T from the partition function can be determined from [34];

$$p_r = \frac{e^{-\beta E_r}}{Z}$$

From the probability and energy, we can obtain the average energy using the energetic population weighted averages [34];

$$\bar{E} = \sum_{r} p_r E_r$$

The implication towards folded proteins can be understood in following manner. Folded proteins are not static (occupancy of a single energy state E_r with associated probability p_r =

1) but can be understood to occupy a finite collection of states energetically proximal to the lowest available energy state.

When discussing disordered structures, a few differences in the treatment of the partition function may be assumed. Compared to the finite number of accessible structures for folded proteins, the distribution of accessible energetic states disordered proteins are sufficiently more diffuse (their average *density of states* (DOS) is much lower). We introduce the concept of DOS to understand IDPs and provide justification for statistical treatment of IDPs.

Previous Characterization of Combinatorial Effects on Protein Structure

Two proteins are considered *circular permutations* of each other if the only difference between two proteins lies in the location of their N and C-termini [35]. For some folded proteins, a degree of similarity between primary sequences that are cyclic permutations of each other display similar three-dimensional (3-D) shape [36]. However, it is unclear if any structural conservation exists due to non-cyclic sequence permutation of proteins. Previous work has shown the large dependence of IDP structure based solely on global factors, particularly PP_{II} conformational propensities and number of peptides N, with a small local charge correction term [25]. Collectively, these results indicate that there is at least some degree of structure that is dictated by global effects, or effects that depend on only combinations of amino acids, regardless of their permutation.

Previous molecular dynamics simulations showed that PP_{II} propensity is a reliable predictor of R_h, which indicates that the structural properties of IDPs rely on a sequence-combination dependent effect (global) as opposed to a sequence-permutation dependent

effect (local) *[25].* Additionally, the thermal denaturation curve for IDPs exhibits noncoopertivity as a linear trend between R_h and temperature is observed for p53(1-93) [37]. It has shown that PP_{II} propensity, measured through GLY → PRO substitutions in p53(1-93) for every proline, does significantly compact R_h, but it is uncertain if the location of these substitutions confer different effects, the hypothesis directly being tested here [30].

Introduction of Methods

The peptide backbone is canonically understood as the source of secondary structure in proteins, making circular dichroism spectroscopy (CD) a suitable method of analysis for understanding the extent of impact on PP_{II} propensity. The spectra of CD reveals information on the secondary structure of proteins through differential absorbance of circularly polarized light, directly measuring in proteins the structure of the amide backbone. In quantitative terms, CD measures the absorbance difference ΔA between left and right-handed circularly polarized light for a given wavelength λ;

$$\Delta A = A_{L\lambda} - A_{R\lambda}$$

Where $A_{R\lambda}$ and $A_{L\lambda}$ represents the absorption of right and left circularly polarized light at wavelength λ. Differential absorbance can be used to derive the molar extinction coefficients ε for right and left circularly polarized light for each wavelength λ and, using Beer's law, extinction coefficients can be converted to concentration normalized absorbance. CD instrumentation measures absorbance directly, using experimentally derived

concentration values and standardized cuvette path length, obtaining a value for Δε, in terms of $M^{-1}cm^{-1}$. Conventionally, CD spectra has units of molar ellipticity, [θ], which is obtained from Δε by;

$$[\theta] = 3298\Delta\varepsilon$$

Yielding the magnitude of the difference between right and left circularly polarized light absorption in concentration normalized units of $deg^{1}cm^{2}dmol^{-1}$. When successive chromophores, in this case, the amide backbone of proteins, are organized into a patterned array, such as a secondary structure motif, the result of CD is a characteristic, reproducible spectra. The PP_{II} helix, previously shown to be an important structural element of disordered proteins, displays a similar spectra to that of random coil proteins, but has a characteristic peak around 220 nm, a wavelength not visible in the random coil spectra [38]. The use of CD on the p53(1-93) substitutions at a variety of temperatures and monitoring the changes in at 221 nm will indicate how the propensity for the PP_{II} helix changes as a result of site-specific substitutions and of directionality.

Size exclusion chromatography (SEC) is a method that can measure the R_h of a given peptide through determination of the length of time needed for a sample to elute through a matrix consisting of homogenously sized beads. By comparing the elution time for the sample to known protein standards, the average sizes can be determined. Due to the lack of stability and the large number of accessible states, the R_h obtained from SEC provides a reproducible metric to empirically compare the effect of p53(1-93) variants and directionality on the average structure of these proteins, which is indicative of their intrinsic conformational propensities [25].

Project Goals

Recent work shows that the prediction of R_h from number of residues and theoretical polyproline-II structure shows good agreement with Pearson's correlation coefficient of 0.93 for 25 proteins experimentally derived R_h through size exclusion chromatography (SEC). This result is not surprising since IDPs lack tertiary stability and thus structural features should be dominated by sequence properties. However, the model implies that the order of sequence does not hold influence in determination of R_h, and thus, sequence-dependent intrinsic conformational propensities.

Ferreon and Hilser performed binding studies on the SH-3 domain of SEM-5 to Sos, a peptide ligand which binds to SH-3 when the peptide is in the PP_{II} conformation [39]. By substituting two different prolines on the 11 peptide ligand to glycine, they were able to determine that a single substitution to glycine was responsible for an average of a 26.7-fold increase in ensemble size over the non-substituted wild-type Sos peptide, as measured with conformational enthalpy of binding using isothermal calorimetry (ITC) . This is consistent with the scale produced by Hilser et. al, using the same SH-3 domain SEM – Sos binding system [28], in which ITC determined that glycine has the lowest fractional propensity of all amino acids for the PP_{II} helix (13% of proline) by measuring the change in enthalpy of binding as a result of different substitutions, causing IDPs to behave more like unbiased random coil ensembles. Because the substitution from proline to glycine should disrupt the PP_{II} helix the most, we designed proline to glycine mutations at different regions of p53(1-

93) should yield the most insight on site-specific effects of intrinsic conformational propensities of IDPs.

The main objective of this project is to understand the degree of local and global sequence regulation in intrinsically disordered proteins. The specific hypothesis tested here is whether there is a significant difference between a collection of seven synthetic variants of the N-terminus of human p53, *p53(1-93)*, which contains 22 proline residues. Tumor suppressor p53, coded by *TP53*, has been widely studied and characterized for its role in disease, and has been referred as the "guardian of the genome", and is the most frequently mutated protein in human cancer [40]. The transactivation domain (TAD) at the amino terminus of p53 is an IDR responsible for binding to cell-signaling related proteins, such as HDM2 (also called MDM2) [41]. Here, the first 93 residues of p53, hereafter refered to as p53(1-93), is used as a model IDP due to 1.) being a relatively well-studied IDP; 2.) functional relevance towards human disease; 3.) highly negative net charge; 4.) high proline content and thus PP_{II} propensity; and 5.) established laboratory protocols for isolation and characterization. Three of the variants have conserved *directionality*, or retention of N- and C-termini; *N11p53*, created by the mutating the first 11 proline residues from the N-terminus to glycine; *M11p53*, created by mutating the middle 11 prolines to glycine; and *C11p53*, created by mutating the last 11 prolines from the N-terminus to glycine. Thus, the three constructed peptide permutations possess the same amino acid constituents as well as directionality, with alterations only in different positions of sequence. Next, the four variants are created similarly to the previous proteins, but with reverse directionality; *p53(93-1),* created from reversing the directionality of p53(1-93), such that the N-terminus and C-terminus are reversed; *RF11p53,* created by reversing the N-terminus and C-terminus of

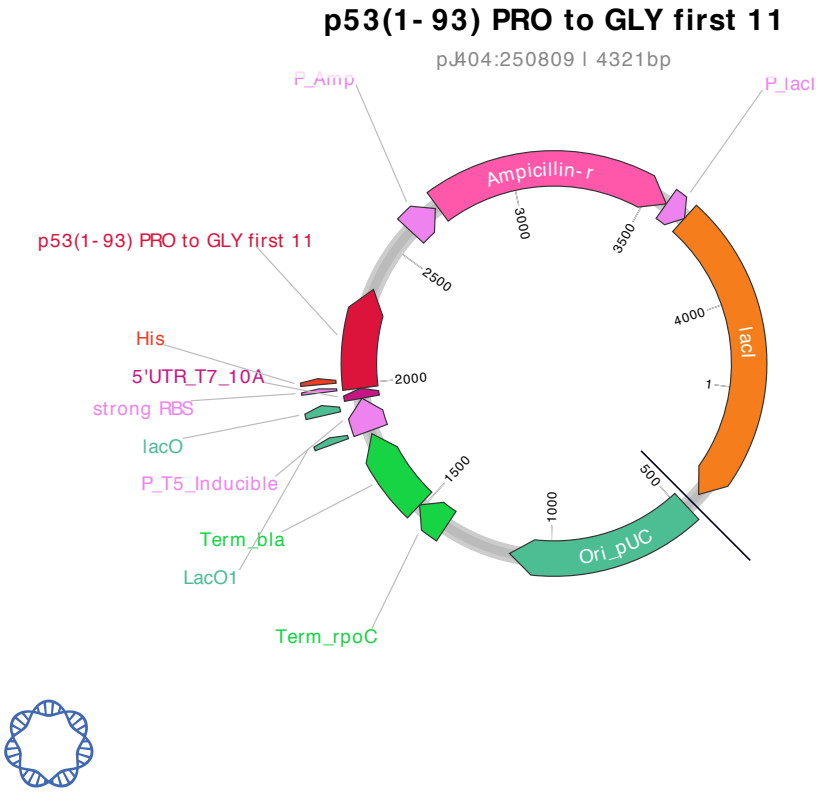
N11p53; *RM11p53*, created by reversing the N-terminus and C-terminus of M11p53; and *RL11p53,* created by reversing the N-terminus and C-terminus of C11p53. These sites are chosen as they convey information on the impact of substitutions in regions with varying charged density; the N11p53/RF11p53 have 13 negatively charged residues between the first and last proline substitutions, and the M11p53/RM11p53 contain 10 negatively charged residues between the first and last proline substituted, while the C11p53/RL11p53 mutants contain only a single glutamate residue between the first and last proline substituted.

The R_h of the isolated and purified peptides is measured using size-exclusion chromatography (SEC). This method is preferred as it is cheaper than other methods of R_h determination (such as sedimentation and NMR spectroscopy) have been used previously for similar peptides, and results have been well characterized in the lab [25, 30, 37]. The null hypothesis is there will be no significant difference between peptide permutations at different locations, as measured with pairwise test on the R_h of each peptide. Secondary structure perturbations are also measured using circular dichroism spectroscopy, as this method has shown success in detecting residual structure in the p53(1-93), and can be used to understand the influence of the PP_{II} helix on compaction of R_h [42]

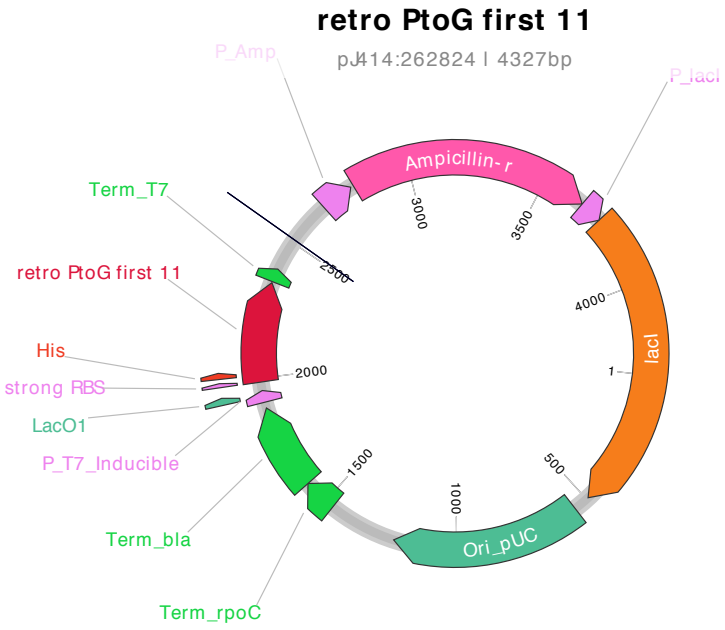
II. MATERIALS AND METHODS

Plasmid and Protein Sequences

The sequence of p53(1-93) was used as the template of this project to understand the effects of different substations on the intrinsic conformational propensities of IDPs, consisting of two charge rich transactivation domain regions (TAD1/TAD2) and a proline rich region (PRR). TAD1 represents residues 1-40, TAD2 represents residues 41-61, and the PRR represents residues from 62-91 [43].



Plasmid map for pJ404 plasmids. The N11p53 variant is illustrated. Figure obtained from ATUM (Newark, CA). Labels: *P_Amp*, a constitutively activated promoter that regulates the Ampicilin resistance gene; *Ampicilin-r*, a gene that confers ampicillin resistance by translating into β-lactamase; *P_lacI*, a constitutively active promoter that regulates the lacI gene; *lacI*, a gene which translates into the lacI protein, which binds *lacO* and *LacO1* and inhibits protein synthesis until induced by IPTG; *Ori_pUC*, which contains the bacterial origin of replication; *Term_rpoC*, contains the transcriptional termination sequence for rpoC, a DNA-dependent RNA polymerase; *Term_bla*, contains the transcription termination sequence for *Ampicillin-r* sequence, which promotes more efficient expression of *Ampicilin-r* gene [44]; *LacO1*, binding site for lac repressor protein; *P_T5_Inducible*, contains the promoter for the T5 bacteriophage RNA polymerase; *lacO*, a secondary site of binding for lac repressor protein; *strong RBS,* a binding sequence which a ribosome binds to the mRNA in translation; *HIS*, a region encoding six histidines attached the protein of interest for purification; *p53(1-93) PRO to GLY first 11,* the region containing the DNA for the protein of interest (either N11p53, M11p53, or C11p53).



Plasmid map for pJ414 plasmids. The RF11p53 variant is illustrated. Figure obtained from ATUM (Newark, CA). Labels: *P_Amp*, a constitutively activated promoter that regulates the Ampicilin resistance gene; *Ampicilin-r*, a gene that confers ampicillin resistance by translating into β-lactamase; *P_lacI*, a constitutively activated promoter that regulates the lacI gene; *lacI*, a gene which translates into the lacI protein, which binds to *LacO1*, inhibiting protein synthesis until induced by IPTG; *Ori_pUC*, which contains the bacterial origin of replication; *Term_rpoC*, contains the transcriptional termination sequence for rpoC, a DNA-dependent RNA polymerase; *Term_bla*, contains the transcription termination sequence for the β-lactamase sequence, which promotes more efficient expression of *Ampicilin-r* gene [44]; *P_T7_Inducible*, contains the promoter for the T7 bacteriophage RNA polymerase; *LacO1*, binding site for lac repressor protein; *strong RBS,* a binding sequence which a ribosome binds to the mRNA in translation; *HIS*, a region encoding six histidines attached the protein of interest for purification; *retro PtoG first 11,* the region containing the DNA for the protein of interest (either p53(1-93), p53(93-1), RF11, RM11, or RL11); *Term_T7*, termination sequence for bacteriophage T7 polymerase.

>wild type p53(93-1)
CCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAG
CGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTTCT
TCTAGTGTAGCCGTAGTTAGCCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCT
GCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGG
GTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCC
CGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGG
TATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAA
ACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGT
GGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCG
GAAGGCGAGAGTAGGGAACTGCCAGGCATCAAACTAAGCAGAAGGCCCCTGACGGATGGCCTTTTTGCGTTTCTACAAACTCTTTCTG
TGTTGTAAAACGACGGCCAGTCTTAAGCTCGGGCCCCCTGGGCGGTTCTGATAACGAGTAATCGTTAATCCGCAAATAACGTAAAAAC
CCGCTTCGGCGGGTTTTTTTATGGGGGGAGTTTAGGGAAAGAGCATTTGTCAGAATATTTAAGGGCGCCTGTCACTTTGCTTGATATA
TGAGAATTATTTAACCTTATAAATGAGAAAAAAGCAACGCACTTTAAATAAGATACGTTGCTTTTTCGATTGATGAACACCTATAATT
AAACTATTCATCTATTATTTATGATTTTTTGTATATACAATATTTCTAGTTTGTTAAAGAGAATTAAGAAAATAAATCTCGAAAATAA
TAAAGGGAAAATCAGTTTTTGATATCAAAATTATACATGTCAACGATAATACAAAATATAATACAAACTATAAGATGTTATCAGTATT
TATTATCATTTAGAATAAATTTTGTGTCGCCCTTCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCC
CTCTAGAAATAATTTTGTTTAACTTTTAAGGAGGTAAAAAATGCGCGGCTCTCATCATCATCACCATCACTCGTCCGGCCTGGTTCCT
CGTGGTAGCATGGAAGAACCGCAATCCGACCCGAGCGTTGAGCCGCCGTTGAGCCAGGAAACCTTCAGCGATCTGTGGAAGCTGCTGC
CGGAGAATAACGTCCTGAGCCCGTTGCCGAGCCAAGCCATGGATGATCTGATGCTGAGCCCGGACGACATCGAGCAGTGGTTTACCGA
AGATCCGGGTCCGGACGAAGCGCCACGTATGCCAGAGGCCGCTCCGCCGGTGGCACCGGCACCGGCAGCGCCGACGCCGGCTGCGCCT
GCGCCAGCGCCGAGCTGGCCGCTGTAACCCCCTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCCCCTGAGA
CGCGTCAATCGAGTTCGTACCTAAGGGCGACACCCCCTAATTAGCCCGGGCGAAAGGCCCAGTCTTTCGACTGAGCCTTTCGTTTTAT
TTGATGCCTGGCAGTTCCCTACTCTCGCATGGGGAGTCCCCACACTACCATCGGCGCTACGGCGTTTCACTTCTGAGTTCGGCATGGG
GTCAGGTGGGACCACCGCGCTACTGCCGCCAGGCAAACAAGGGGTGTTATGAGCCATATTCAGGTATAAATGGGCTCGCGATAATGTT
CAGAATTGGTTAATTGGTTGTAACACTGACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAA
CCCTGATAAATGCTTCAATAATATTGAAAAAGGAAGAATATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCAT
TTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGA
ACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGT
GGCGCGGTATTATCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACTCAC
CAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAA
CTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGG
GAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCGATGGCAACAACGTTGCGCAAACTATTAA
CTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGC
CCTTCCGGCTGGCTGGTTTATTGCTGATAAATCCGGAGCCGGTGAGCGTGGTTCTCGCGGTATCATCGCAGCGCTGGGGCCAGATGGT
AAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCT
CACTGATTAAGCATTGGTAAGCGGCGCGCCATCGAATGGCGCAAAACCTTTCGCGGTATGGCATGATAGCGCCCGGAAGAGAGTCAAT
TCAGGGTGGTGAATATGAAACCAGTAACGTTATACGATGTCGCAGAGTATGCCGGTGTCTCTTATCAGACCGTTTCCCGCGTGGTGAA
CCAGGCCAGCCACGTTTCTGCGAAAACGCGGGAAAAAGTGGAAGCGGCGATGGCGGAGCTGAATTACATTCCCAACCGCGTGGCACAA
CAACTGGCGGGCAAACAGTCGTTGCTGATTGGCGTTGCCACCTCCAGTCTGGCCCTGCACGCGCCGTCGCAAATTGTCGCGGCGATTA
AATCTCGCGCCGATCAACTGGGTGCCAGCGTGGTGGTGTCGATGGTAGAACGAAGCGGCGTCGAAGCCTGTAAAGCGGCGGTGCACAA
TCTTCTCGCGCAACGCGTCAGTGGGCTGATCATTAACTATCCGCTGGATGACCAGGATGCCATTGCTGTGGAAGCTGCCTGCACTAAT
GTTCCGGCGTTATTTCTTGATGTCTCTGACCAGACACCCATCAACAGTATTATTTTCTCCCATGAGGACGGTACGCGACTGGGCGTGG
AGCATCTGGTCGCATTGGGTCACCAGCAAATCGCGCTGTTAGCGGGCCCATTAAGTTCTGTCTCGGCGCGTCTGCGTCTGGCTGGCTG
GCATAAATATCTCACTCGCAATCAAATTCAGCCGATAGCGGAACGGGAAGGCGACTGGAGTGCCATGTCCGGTTTTCAACAAACCATG
CAAATGCTGAATGAGGGCATCGTTCCCACTGCGATGCTGGTTGCCAACGATCAGATGGCGCTGGGCGCAATGCGCGCCATTACCGAGT
CCGGGCTGCGCGTTGGTGCGGATATCTCGGTAGTGGGATACGACGATACCGAAGATAGCTCATGTTATATCCCGCCGTTAACCACCAT
CAAACAGGATTTTCGCCTGCTGGGGCAAACCAGCGTGGACCGCTTGCTGCAACTCTCTCAGGGCCAGGCGGTGAAGGGCAATCAGCTG
TTGCCAGTCTCACTGGTGAAAAGAAAAACCACCCTGGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAATGC
AGCTGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAGTGACTCATGACCAAAATCCCTTAACGTGAGTTACGCGCGCGTCGTTCCAC
TGAGCGTCAGAC

FIG. 1. **FASTA sequence of pJ414 plasmid containing p53(1-93).** Sequences are highlighted according to the following legend: *Ori_pUC*, *Term_rpoC*, *term_bla*, *P_T7_Inducible*, *T7/LacO1*, *RBS*, *HIS*, *p53(93-1)*, *Term_t7*, *AMP_p*, *Ampicilin-r*, *P_lacI*, *lacI*. This plasmid contains 4324 total base pairs

>retro wild type p53(93-1)

CCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAG
CGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTTCT
TCTAGTGTAGCCGTAGTTAGCCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCT
GCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGG
GTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCC
CGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGG
TATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAA
ACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGT
GGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCG
GAAGGCGAGAGTAGGGAACTGCCAGGCATCAAACTAAGCAGAAGGCCCCTGACGGATGGCCTTTTTGCGTTTCTACAAACTCTTTCTG
TGTTGTAAAACGACGGCCAGTCTTAAGCTCGGGCCCCCTGGGCGGTTCTGATAACGAGTAATCGTTAATCCGCAAATAACGTAAAAAC
CCGCTTCGGCGGGTTTTTTTATGGGGGGAGTTTAGGGAAAGAGCATTTGTCAGAATATTTAAGGGCGCCTGTCACTTTGCTTGATATA
TGAGAATTATTTAACCTTATAAATGAGAAAAAAGCAACGCACTTTAAATAAGATACGTTGCTTTTTCGATTGATGAACACCTATAATT
AAACTATTCATCTATTATTTATGATTTTTTGTATATACAATATTTCTAGTTTGTTAAAGAGAATTAAGAAAATAAATCTCGAAAATAA
TAAAGGGAAAATCAGTTTTTGATATCAAAATTATACATGTCAACGATAATACAAAATATAATACAAACTATAAGATGTTATCAGTATT
TATTATCATTTAGAATAAATTTTGTGTCGCCCTTCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCC
CTCTAGAAATAATTTTGTTTAACTTTTAAGGAGGTAAAAAATGCGCGGCTCTCACCATCACCATCATCACTCGTCCGGCCTTGTCCCA
CGTGGTAGCTTGCCGTGGTCCCCAGCCCCTGCTCCGGCAGCGCCGACGCCGGCAGCCCCGGCGCCGGCTGTTCCGCCAGCGGCGGAGC
CGATGCGTCCGGCAGAGGATCCGGGTCCGGACGAAACCTTTTGGCAAGAAATCGACGATCCGAGCCTGATGCTGGATGATATGGCGCA
GAGCCCGCTGCCGAGCCTGGTGAATAACGAACCGCTGCTGAAATGGCTGGACAGCTTCACCGAGCAAAGCCTGCCGCCTGAAGTTAGC
CCGGACAGCCAGCCGGAAGAAATGTAACCCCCTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCCCCTGAGA
CGCGTCAATCGAGTTCGTACCTAAGGGCGACACCCCCTAATTAGCCCGGGCGAAAGGCCCAGTCTTTCGACTGAGCCTTTCGTTTTAT
TTGATGCCTGGCAGTTCCCTACTCTCGCATGGGGAGTCCCCACACTACCATCGGCGCTACGGCGTTTCACTTCTGAGTTCGGCATGGG
GTCAGGTGGGACCACCGCGCTACTGCCGCCAGGCAAACAAGGGGTGTTATGAGCCATATTCAGGTATAAATGGGCTCGCGATAATGTT
CAGAATTGGTTAATTGGTTGTAACACTGACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAA
CCCTGATAAATGCTTCAATAATATTGAAAAAGGAAGAATATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCAT
TTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGA
ACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGT
GGCGCGGTATTATCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACTCAC
CAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAA
CTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGG
GAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCGATGGCAACAACGTTGCGCAAACTATTAA
CTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGC
CCTTCCGGCTGGCTGGTTTATTGCTGATAAATCCGGAGCCGGTGAGCGTGGTTCTCGCGGTATCATCGCAGCGCTGGGGCCAGATGGT
AAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCT
CACTGATTAAGCATTGGTAAGCGGCGCGCCATCGAATGGCGCAAAACCTTTCGCGGTATGGCATGATAGCGCCCGGAAGAGAGTCAAT
TCAGGGTGGTGAATATGAAACCAGTAACGTTATACGATGTCGCAGAGTATGCCGGTGTCTCTTATCAGACCGTTTCCCGCGTGGTGAA
CCAGGCCAGCCACGTTTCTGCGAAAACGCGGGAAAAAGTGGAAGCGGCGATGGCGGAGCTGAATTACATTCCCAACCGCGTGGCACAA
CAACTGGCGGGCAAACAGTCGTTGCTGATTGGCGTTGCCACCTCCAGTCTGGCCCTGCACGCGCCGTCGCAAATTGTCGCGGCGATTA
AATCTCGCGCCGATCAACTGGGTGCCAGCGTGGTGGTGTCGATGGTAGAACGAAGCGGCGTCGAAGCCTGTAAAGCGGCGGTGCACAA
TCTTCTCGCGCAACGCGTCAGTGGGCTGATCATTAACTATCCGCTGGATGACCAGGATGCCATTGCTGTGGAAGCTGCCTGCACTAAT
GTTCCGGCGTTATTTCTTGATGTCTCTGACCAGACACCCATCAACAGTATTATTTTCTCCCATGAGGACGGTACGCGACTGGGCGTGG
AGCATCTGGTCGCATTGGGTCACCAGCAAATCGCGCTGTTAGCGGGCCCATTAAGTTCTGTCTCGGCGCGTCTGCGTCTGGCTGGCTG
GCATAAATATCTCACTCGCAATCAAATTCAGCCGATAGCGGAACGGGAAGGCGACTGGAGTGCCATGTCCGGTTTTCAACAAACCATG
CAAATGCTGAATGAGGGCATCGTTCCCACTGCGATGCTGGTTGCCAACGATCAGATGGCGCTGGGCGCAATGCGCGCCATTACCGAGT
CCGGGCTGCGCGTTGGTGCGGATATCTCGGTAGTGGGATACGACGATACCGAAGATAGCTCATGTTATATCCCGCCGTTAACCACCAT
CAAACAGGATTTTCGCCTGCTGGGGCAAACCAGCGTGGACCGCTTGCTGCAACTCTCTCAGGGCCAGGCGGTGAAGGGCAATCAGCTG
TTGCCAGTCTCACTGGTGAAAAGAAAAACCACCCTGGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAATGC
AGCTGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAGTGACTCATGACCAAAATCCCTTAACGTGAGTTACGCGCGCGTCGTTCCAC
TGAGCGTCAGAC

FIG. 2. **FASTA sequence of pJ414 plasmid containing p53(93-1).** Sequences are highlighted according to the following legend: *Ori_pUC*, *Term_rpoC*, *term_bla*, *P_T7_Inducible*, *T7/LacO1*, *RBS*, *HIS*, *p53(93-1)*, *Term_t7*, *AMP_p*, *Ampicilin-r*, *P_lacI*, *lacI*. This plasmid contains 4324 total base pairs.

>retro PtoG first 11 (RF11p53)
CCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAG
CGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTTCT
TCTAGTGTAGCCGTAGTTAGCCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCT
GCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGG
GTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCC
CGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGG
TATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAA
ACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGT
GGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCG
GAAGGCGAGAGTAGGGAACTGCCAGGCATCAAACTAAGCAGAAGGCCCCTGACGGATGGCCTTTTTGCGTTTCTACAAACTCTTTCTG
TGTTGTAAAACGACGGCCAGTCTTAAGCTCGGGCCCCCTGGGCGGTTCTGATAACGAGTAATCGTTAATCCGCAAATAACGTAAAAAC
CCGCTTCGGCGGGTTTTTTTTATGGGGGGAGTTTAGGGAAAGAGCATTTGTCAGAATATTTAAGGGCGCCTGTCACTTTGCTTGATATA
TGAGAATTATTTAACCTTATAAATGAGAAAAAAGCAACGCACTTTAAATAAGATACGTTGCTTTTTCGATTGATGAACACCTATAATT
AAACTATTCATCTATTATTTATGATTTTTTGTATATACAATATTTCTAGTTTGTTAAAGAGAATTAAGAAAATAAATCTCGAAAATAA
TAAAGGGAAAATCAGTTTTTGATATCAAAATTATACATGTCAACGATAATACAAAATATAATACAAACTATAAGATGTTATCAGTATT
TATTATCATTTAGAATAAATTTTGTGTCGCCCTTCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCC
CTCTAGAAATAATTTTGTTTAACTTTTAAGGAGGTAAAAAATGCGCGGCTCTCATCATCATCACCATCACAGCTCCGGTCTGGTCCCA
CGTGGCTCTCTGCCGTGGAGCCCTGCGCCGGCACCGGCTGCACCGACGCCGGCGGCACCGGCGCCGGCTGTTCCGCCGGCGGCCGAGC
CGATGCGTGGTGCCGAGGATGGTGGCGGCGACGAAACCTTCTGGCAAGAAATCGACGACGGTAGCCTGATGCTGGATGATATGGCGCA
GAGCGGCCTGGGTAGCTTGGTGAATAACGAGGGCTTGCTGAAATGGCTGGATAGCTTTACCGAGCAATCGCTGGGTGGCGAAGTTAGC
GGTGACAGCCAGGGTGAAGAAATGTAATGACCCCCTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCCCCTG
AGACGCGTCAATCGAGTTCGTACCTAAGGGCGACACCCCCTAATTAGCCCGGGCGAAAGGCCCAGTCTTTCGACTGAGCCTTTCGTTT
TATTTGATGCCTGGCAGTTCCCTACTCTCGCATGGGGAGTCCCCACACTACCATCGGCGCTACGGCGTTTCACTTCTGAGTTCGGCAT
GGGGTCAGGTGGGACCACCGCGCTACTGCCGCCAGGCAAACAAGGGGTGTTATGAGCCATATTCAGGTATAAATGGGCTCGCGATAAT
GTTCAGAATTGGTTAATTGGTTGTAACACTGACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAA
TAACCCTGATAAATGCTTCAATAATATTGAAAAAGGAAGAATATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGG
CATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACAT
CGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTA
TGTGGCGCGGTATTATCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACT
CACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGC
CAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGT
TGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCGATGGCAACAACGTTGCGCAAACTAT
TAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTC
GGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCCGGAGCCGGTGAGCGTGGTTCTCGCGGTATCATCGCAGCGCTGGGGCCAGAT
GGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTG
CCTCACTGATTAAGCATTGGTAAGCGGCGCGCCATCGAATGGCGCAAAACCTTTCGCGGTATGGCATGATAGCGCCCGGAAGAGAGTC
AATTCAGGGTGGTGAATATGAAACCAGTAACGTTATACGATGTCGCAGAGTATGCCGGTGTCTCTTATCAGACCGTTTCCCGCGTGGT
GAACCAGGCCAGCCACGTTTCTGCGAAAACGCGGGAAAAAGTGGAAGCGGCGATGGCGGAGCTGAATTACATTCCCAACCGCGTGGCA
CAACAACTGGCGGGCAAACAGTCGTTGCTGATTGGCGTTGCCACCTCCAGTCTGGCCCTGCACGCGCCGTCGCAAATTGTCGCGGCGA
TTAAATCTCGCGCCGATCAACTGGGTGCCAGCGTGGTGGTGTCGATGGTAGAACGAAGCGGCGTCGAAGCCTGTAAAGCGGCGGTGCA
CAATCTTCTCGCGCAACGCGTCAGTGGGCTGATCATTAACTATCCGCTGGATGACCAGGATGCCATTGCTGTGGAAGCTGCCTGCACT
AATGTTCCGGCGTTATTTCTTGATGTCTCTGACCAGACACCCATCAACAGTATTATTTTCTCCCATGAGGACGGTACGCGACTGGGCG
TGGAGCATCTGGTCGCATTGGGTCACCAGCAAATCGCGCTGTTAGCGGGCCCATTAAGTTCTGTCTCGGCGCGTCTGCGTCTGGCTGG
CTGGCATAAATATCTCACTCGCAATCAAATTCAGCCGATAGCGGAACGGGAAGGCGACTGGAGTGCCATGTCCGGTTTTCAACAAACC
ATGCAAATGCTGAATGAGGGCATCGTTCCCACTGCGATGCTGGTTGCCAACGATCAGATGGCGCTGGGCGCAATGCGCGCCATTACCG
AGTCCGGGCTGCGCGTTGGTGCGGATATCTCGGTAGTGGGATACGACGATACCGAAGATAGCTCATGTTATATCCCGCCGTTAACCAC
CATCAAACAGGATTTTCGCCTGCTGGGGCAAACCAGCGTGGACCGCTTGCTGCAACTCTCTCAGGGCCAGGCGGTGAAGGGCAATCAG
CTGTTGCCAGTCTCACTGGTGAAAAGAAAAACCACCCTGGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAA
TGCAGCTGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAGTGACTCATGACCAAAATCCCTTAACGTGAGTTACGCGCGCGTCGTTC
CACTGAGCGTCAGAC

FIG. 3. **FASTA sequence of pJ414 plasmid containing RF11p53.** Sequences are highlighted according to the following legend: *Ori_pUC*, *Term_rpoC*, *term_bla*, *P_T7_Inducible*, *T7/LacO1*, *RBS*, *HIS*, *p53(93-1)*, *Term_t7*, *AMP_p*, *Ampicilin-r*, *P_lacI*, *lacI*. This plasmid contains 4327 total base pairs.

>retro PtoG middle 11 (RM11p53)

CCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAG
CGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTTCT
TCTAGTGTAGCCGTAGTTAGCCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCT
GCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGG
GTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCC
CGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGG
TATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAA
ACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGT
GGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCG
GAAGGCGAGAGTAGGGAACTGCCAGGCATCAAACTAAGCAGAAGGCCCCTGACGGATGGCCTTTTTGCGTTTCTACAAACTCTTTCTG
TGTTGTAAAACGACGGCCAGTCTTAAGCTCGGGCCCCCTGGGCGGTTCTGATAACGAGTAATCGTTAATCCGCAAATAACGTAAAAAC
CCGCTTCGGCGGGTTTTTTTTATGGGGGGAGTTTAGGGAAAGAGCATTTGTCAGAATATTTAAGGGCGCCTGTCACTTTGCTTGATATA
TGAGAATTATTTAACCTTATAAATGAGAAAAAAGCAACGCACTTTAAATAAGATACGTTGCTTTTTCGATTGATGAACACCTATAATT
AAACTATTCATCTATTATTTATGATTTTTTGTATATACAATTTCTAGTTTGTTAAAGAGAATTAAGAAAATAAATCTCGAAAATAA
TAAAGGGAAAATCAGTTTTTGATATCAAAATTATACATGTCAACGATAATACAAAATATAATACAAACTATAAGATGTTATCAGTATT
TATTATCATTTAGAATAAATTTTGTGTCGCCCTTCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCC
CTCTAGAAATAATTTTGTTTAACTTTTAAGGAGGTAAAAAATGCGCGGTTCTCATCATCACCATCATCACAGCAGCGGTCTGGTCCCT
CGTGGCAGCTTGGGCTGGAGCGGTGCAGGCGCGGGCGCGGCTGGTACCGGTGCGGCGGGTGCGGGTGCCGTTGGTGGCGCAGCTGAGG
GCATGCGTCCGGCCGAGGATCCGGGTCCGGACGAAACGTTTTGGCAAGAAATCGACGATCCGAGCCTGATGCTGGACGACATGGCACA
GTCGCCGCTGCCGAGCCTGGTGAACAATGAACCGTTGCTGAAATGGCTGGATAGCTTCACCGAGCAGTCCCTGCCGCCAGAAGTGAGC
CCGGATTCCCAACCGGAAGAAATGTGATAACCCCCTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCCCCTG
AGACGCGTCAATCGAGTTCGTACCTAAGGGCGACACCCCCTAATTAGCCCGGGCGAAAGGCCCAGTCTTTCGACTGAGCCTTTCGTTT
TATTTGATGCCTGGCAGTTCCCTACTCTCGCATGGGGAGTCCCCACACTACCATCGGCGCTACGGCGTTTCACTTCTGAGTTCGGCAT
GGGGTCAGGTGGGACCACCGCGCTACTGCCGCCAGGCAAACAAGGGGTGTTATGAGCCATATTCAGGTATAAATGGGCTCGCGATAAT
GTTCAGAATTGGTTAATTGGTTGTAACACTGACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAA
TAACCCTGATAAATGCTTCAATAATATTGAAAAAGGAAGAATATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGG
CATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACAT
CGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTA
TGTGGCGCGGTATTATCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACT
CACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGC
CAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGT
TGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCGATGGCAACAACGTTGCGCAAACTAT
TAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTC
GGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCCGGAGCCGGTGAGCGTGGTTCTCGCGGTATCATCGCAGCGCTGGGGCCAGAT
GGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTG
CCTCACTGATTAAGCATTGGTAAGCGGCGCGCCATCGAATGGCGCAAAACCTTTCGCGGTATGGCATGATAGCGCCCGGAAGAGAGTC
AATTCAGGGTGGTGAATATGAAACCAGTAACGTTATACGATGTCGCAGAGTATGCCGGTGTCTCTTATCAGACCGTTTCCCGCGTGGT
GAACCAGGCCAGCCACGTTTCTGCGAAAACGCGGGAAAAAGTGGAAGCGGCGATGGCGGAGCTGAATTACATTCCCAACCGCGTGGCA
CAACAACTGGCGGGCAAACAGTCGTTGCTGATTGGCGTTGCCACCTCCAGTCTGGCCCTGCACGCGCCGTCGCAAATTGTCGCGGCGA
TTAAATCTCGCGCCGATCAACTGGGTGCCAGCGTGGTGGTGTCGATGGTAGAACGAAGCGGCGTCGAAGCCTGTAAAGCGGCGGTGCA
CAATCTTCTCGCGCAACGCGTCAGTGGGCTGATCATTAACTATCCGCTGGATGACCAGGATGCCATTGCTGTGGAAGCTGCCTGCACT
AATGTTCCGGCGTTATTTCTTGATGTCTCTGACCAGACACCCATCAACAGTATTATTTTCTCCCATGAGGACGGTACGCGACTGGGCG
TGGAGCATCTGGTCGCATTGGGTCACCAGCAAATCGCGCTGTTAGCGGGCCCATTAAGTTCTGTCTCGGCGCGTCTGCGTCTGGCTGG
CTGGCATAAATATCTCACTCGCAATCAAATTCAGCCGATAGCGGAACGGGAAGGCGACTGGAGTGCCATGTCCGGTTTTCAACAAACC
ATGCAAATGCTGAATGAGGGCATCGTTCCCACTGCGATGCTGGTTGCCAACGATCAGATGGCGCTGGGCGCAATGCGCGCCATTACCG
AGTCCGGGCTGCGCGTTGGTGCGGATATCTCGGTAGTGGGATACGACGATACCGAAGATAGCTCATGTTATATCCCGCCGTTAACCAC
CATCAAACAGGATTTTCGCCTGCTGGGGCAAACCAGCGTGGACCGCTTGCTGCAACTCTCTCAGGGCCAGGCGGTGAAGGGCAATCAG
CTGTTGCCAGTCTCACTGGTGAAAAGAAAAACCACCCTGGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAA
TGCAGCTGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAGTGACTCATGACCAAAATCCCTTAACGTGAGTTACGCGCGCGTCGTTC
CACTGAGCGTCAGAC

FIG. 4. **FASTA sequence of pJ414 plasmid containing RM11p53.** Sequences are highlighted according to the following legend: *Ori_pUC*, *Term_rpoC*, *term_bla*, *P_T7_Inducible*, *T7/LacO1*, *RBS*, *HIS*, *p53(93-1)*, *Term_t7*, *AMP_p*, *Ampicilin-r*, *P_lacI*, *lacI*. This plasmid contains 4327 total base pairs.

>retro PtoG last 11 (RL11p53)
CCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAG
CGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTTCT
TCTAGTGTAGCCGTAGTTAGCCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCT
GCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGG
GTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCC
CGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGG
TATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAA
ACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGT
GGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCG
GAAGGCGAGAGTAGGGAACTGCCAGGCATCAAACTAAGCAGAAGGCCCCTGACGGATGGCCTTTTTGCGTTTCTACAAACTCTTTCTG
TGTTGTAAAACGACGGCCAGTCTTAAGCTCGGGCCCCCTGGGCGGTTCTGATAACGAGTAATCGTTAATCCGCAAATAACGTAAAAAC
CCGCTTCGGCGGGTTTTTTTTATGGGGGGAGTTTAGGGAAAGAGCATTTGTCAGAATATTTAAGGGCGCCTGTCACTTTGCTTGATATA
TGAGAATTATTTAACCTTATAAATGAGAAAAAAGCAACGCACTTTAAATAAGATACGTTGCTTTTTCGATTGATGAACACCTATAATT
AAACTATTCATCTATTATTTATGATTTTTTGTATATACAATTTCTAGTTTGTTAAAGAGAATTAAGAAAATAAATCTCGAAAATAA
TAAAGGGAAAATCAGTTTTTGATATCAAAATTATACATGTCAACGATAATACAAAATATAATACAAACTATAAGATGTTATCAGTATT
TATTATCATTTAGAATAAATTTTGTGTCGCCCTTCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCC
CTCTAGAAATAATTTTGTTTAACTTTTAAGGAGGTAAAAAATGCGCGGTTCTCATCATCACCATCATCACAGCAGCGGTCTGGTCCCT
CGTGGCAGCTTGGGCTGGAGCGGTGCAGGCGCGGGCGCGGCTGGTACCGGTGCGGCGGGTGCGGGTGCCGTTGGTGGCGCAGCTGAGG
GCATGCGTCCGGCCGAGGATCCGGGTCCGGACGAAACGTTTTGGCAAGAAATCGACGATCCGAGCCTGATGCTGGACGACATGGCACA
GTCGCCGCTGCCGAGCCTGGTGAACAATGAACCGTTGCTGAAATGGCTGGATAGCTTCACCGAGCAGTCCCTGCCGCCAGAAGTGAGC
CCGGATTCCCAACCGGAAGAAATGTGATAACCCCCTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCCCCTG
AGACGCGTCAATCGAGTTCGTACCTAAGGGCGACACCCCCTAATTAGCCCGGGCGAAAGGCCCAGTCTTTCGACTGAGCCTTTCGTTT
TATTTGATGCCTGGCAGTTCCCTACTCTCGCATGGGGAGTCCCCACACTACCATCGGCGCTACGGCGTTTCACTTCTGAGTTCGGCAT
GGGGTCAGGTGGGACCACCGCGCTACTGCCGCCAGGCAAACAAGGGGTGTTATGAGCCATATTCAGGTATAAATGGGCTCGCGATAAT
GTTCAGAATTGGTTAATTGGTTGTAACACTGACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAA
TAACCCTGATAAATGCTTCAATAATATTGAAAAAGGAAGAATATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGG
CATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACAT
CGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTA
TGTGGCGCGGTATTATCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACT
CACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGC
CAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGT
TGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCGATGGCAACAACGTTGCGCAAACTAT
TAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTC
GGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCCGGAGCCGGTGAGCGTGGTTCTCGCGGTATCATCGCAGCGCTGGGGCCAGAT
GGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTG
CCTCACTGATTAAGCATTGGTAAGCGGCGCGCCATCGAATGGCGCAAAACCTTTCGCGGTATGGCATGATAGCGCCCGGAAGAGAGTC
AATTCAGGGTGGTGAATATGAAACCAGTAACGTTATACGATGTCGCAGAGTATGCCGGTGTCTCTTATCAGACCGTTTCCCGCGTGGT
GAACCAGGCCAGCCACGTTTCTGCGAAAACGCGGGAAAAAGTGGAAGCGGCGATGGCGGAGCTGAATTACATTCCCAACCGCGTGGCA
CAACAACTGGCGGGCAAACAGTCGTTGCTGATTGGCGTTGCCACCTCCAGTCTGGCCCTGCACGCGCCGTCGCAAATTGTCGCGGCGA
TTAAATCTCGCGCCGATCAACTGGGTGCCAGCGTGGTGGTGTCGATGGTAGAACGAAGCGGCGTCGAAGCCTGTAAAGCGGCGGTGCA
CAATCTTCTCGCGCAACGCGTCAGTGGGCTGATCATTAACTATCCGCTGGATGACCAGGATGCCATTGCTGTGGAAGCTGCCTGCACT
AATGTTCCGGCGTTATTTCTTGATGTCTCTGACCAGACACCCATCAACAGTATTATTTTCTCCCATGAGGACGGTACGCGACTGGGCG
TGGAGCATCTGGTCGCATTGGGTCACCAGCAAATCGCGCTGTTAGCGGGCCCATTAAGTTCTGTCTCGGCGCGTCTGCGTCTGGCTGG
CTGGCATAAATATCTCACTCGCAATCAAATTCAGCCGATAGCGGAACGGGAAGGCGACTGGAGTGCCATGTCCGGTTTTCAACAAACC
ATGCAAATGCTGAATGAGGGCATCGTTCCCACTGCGATGCTGGTTGCCAACGATCAGATGGCGCTGGGCGCAATGCGCGCCATTACCG
AGTCCGGGCTGCGCGTTGGTGCGGATATCTCGGTAGTGGGATACGACGATACCGAAGATAGCTCATGTTATATCCCGCCGTTAACCAC
CATCAAACAGGATTTTCGCCTGCTGGGGCAAACCAGCGTGGACCGCTTGCTGCAACTCTCTCAGGGCCAGGCGGTGAAGGGCAATCAG
CTGTTGCCAGTCTCACTGGTGAAAAGAAAAACCACCCTGGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAA
TGCAGCTGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAGTGACTCATGACCAAAATCCCTTAACGTGAGTTACGCGCGCGTCGTTC
CACTGAGCGTCAGAC

FIG. 5. **FASTA sequence of pJ414 plasmid containing RL11p53.** Sequences are highlighted according to the following legend: *Ori_pUC*, *Term_rpoC*, *term_bla*, *P_T7_Inducible*, *T7/LacO1*, *RBS*, *HIS*, *p53(93-1)*, *Term_t7*, *AMP_p*, *Ampicilin-r*, *P_lacI*, *lacI*. This plasmid contains 4327 total base pairs.

>p53(1-93) PRO to GLY first 11
CCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAG
CGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTTCT
TCTAGTGTAGCCGTAGTTAGCCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCT
GCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGG
GTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCC
CGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGG
TATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAA
ACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGT
GGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCG
GAAGGCGAGAGTAGGGAACTGCCAGGCATCAAACTAAGCAGAAGGCCCCTGACGGATGGCCTTTTTGCGTTTCTACAAACTCTTTCTG
TGTTGTAAAACGACGGCCAGTCTTAAGCTCGGGCCCCCTGGGCGGTTCTGATAACGAGTAATCGTTAATCCGCAAATAACGTAAAAAC
CCGCTTCGGCGGGTTTTTTTTATGGGGGGAGTTTAGGGAAAGAGCATTTGTCAGAATATTTAAGGGCGCCTGTCACTTTGCTTGATATA
TGAGAATTATTTAACCTTATAAATGAGAAAAAAGCAACGCACTTTAAATAAGATACGTTGCTTTTTCGATTGATGAACACCTATAATT
AAACTATTCATCTATTATTTATGATTTTTTGTATATACAATTTCTAGTTTGTTAAAGAGAATTAAGAAAATAAATCTCGAAAATAA
TAAAGGGAAAATCAGTTTTTGATATCAAAATTATACATGTCAACGATAATACAAAATATAATACAAACTATAAGATGTTATCAGTATT
TATTATCATTTAGAATAAATTTTGTGTCGCCCTTAATTGTGAGCGGATAACAATTACGAGCTTCATGCACAGTGAAATCATGAAAAAT
TTATTTGCTTTGTGAGCGGATAACAATTATAATATGTGGAATTGTGAGCGCTCACAATTCCACAACGGTTTCCCTCTAGAAATAATTT
TGTTTAACTTTTAAGGAGGTAAAAAATGCGCGGTTCTCATCATCATCACCATCACAGCTCTGGTCTGGTCCCTCGTGGTAGCATGGAA
GAGGGCCAGAGCGACGGTTCCGTTGAGGGTGGCTTGAGCCAAGAAACGTTTAGCGATCTGTGGAAACTGCTGGGTGAGAATAACGTGC
TGTCGGGTCTGGGTAGCCAAGCCATGGATGACCTGATGCTGAGCGGCGACGACATCGAGCAGTGGTTCACCGAAGATGGCGGTGGCGA
TGAAGCTGGCCGTATGCCGGAAGCGGCACCGCCGGTGGCGCCAGCGCCGGCGGCGCCGACCCCGGCAGCCCCGGCACCGGCTCCGAGC
TGGCCGCTGTGATAACCCCAAGGGCGACACCCCATAATTAGCCCGGGCGAAAGGCCCAGTCTTTCGACTGAGCCTTTCGTTTTATTTG
ATGCCTGGCAGTTCCCTACTCTCGCATGGGGAGTCCCCACACTACCATCGGCGCTACGGCGTTTCACTTCTGAGTTCGGCATGGGGTC
AGGTGGGACCACCGCGCTACTGCCGCCAGGCAAACAAGGGGTGTTATGAGCCATATTCAGGTATAAATGGGCTCGCGATAATGTTCAG
AATTGGTTAATTGGTTGTAACACTGACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAACCC
TGATAAATGCTTCAATAATATTGAAAAAGGAAGAATATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTT
GCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACT
GGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGC
GCGGTATTATCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACTCACCAG
TCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTT
ACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAA
CCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCGATGGCAACAACGTTGCGCAAACTATTAACTG
GCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCT
TCCGGCTGGCTGGTTTATTGCTGATAAATCCGGAGCCGGTGAGCGTGGTTCTCGCGGTATCATCGCAGCGCTGGGGCCAGATGGTAAG
CCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCAC
TGATTAAGCATTGGTAAGCGGCGCGCCATCGAATGGCGCAAAACCTTTCGCGGTATGGCATGATAGCGCCCGGAAGAGAGTCAATTCA
GGGTGGTGAATATGAAACCAGTAACGTTATACGATGTCGCAGAGTATGCCGGTGTCTCTTATCAGACCGTTTCCCGCGTGGTGAACCA
GGCCAGCCACGTTTCTGCGAAAACGCGGGAAAAAGTGGAAGCGGCGATGGCGGAGCTGAATTACATTCCCAACCGCGTGGCACAACAA
CTGGCGGGCAAACAGTCGTTGCTGATTGGCGTTGCCACCTCCAGTCTGGCCCTGCACGCGCCGTCGCAAATTGTCGCGGCGATTAAAT
CTCGCGCCGATCAACTGGGTGCCAGCGTGGTGGTGTCGATGGTAGAACGAAGCGGCGTCGAAGCCTGTAAAGCGGCGGTGCACAATCT
TCTCGCGCAACGCGTCAGTGGGCTGATCATTAACTATCCGCTGGATGACCAGGATGCCATTGCTGTGGAAGCTGCCTGCACTAATGTT
CCGGCGTTATTTCTTGATGTCTCTGACCAGACACCCATCAACAGTATTATTTTCTCCCATGAGGACGGTACGCGACTGGGCGTGGAGC
ATCTGGTCGCATTGGGTCACCAGCAAATCGCGCTGTTAGCGGGCCCATTAAGTTCTGTCTCGGCGCGTCTGCGTCTGGCTGGCTGGCA
TAAATATCTCACTCGCAATCAAATTCAGCCGATAGCGGAACGGGAAGGCGACTGGAGTGCCATGTCCGGTTTTCAACAAACCATGCAA
ATGCTGAATGAGGGCATCGTTCCCACTGCGATGCTGGTTGCCAACGATCAGATGGCGCTGGGCGCAATGCGCGCCATTACCGAGTCCG
GGCTGCGCGTTGGTGCGGATATCTCGGTAGTGGGATACGACGATACCGAAGATAGCTCATGTTATATCCCGCCGTTAACCACCATCAA
ACAGGATTTTCGCCTGCTGGGGCAAACCAGCGTGGACCGCTTGCTGCAACTCTCTCAGGGCCAGGCGGTGAAGGGCAATCAGCTGTTG
CCAGTCTCACTGGTGAAAAGAAAAACCACCCTGGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAATGCAGC
TGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAGTGACTCATGACCAAAATCCCTTAACGTGAGTTACGCGCGCGTCGTTCCACTGA
GCGTCAGAC

FIG. 6. **FASTA sequence of pJ404 plasmid containing N11p53.** Sequences are highlighted according to the following legend: *Ori_puC*, *term_rpoc*, *Term_bla*, *LacO1/P_T5_Inducible*, *P_T5_inducible*, *lacO/P_T5_inducible*, *5'UTR_T7_10A*, *RBS*, *HIS*, *AMP_p*, *Ampicilin-r*, *P_lacI*, *lacI*. This plasmid contains 4321 total base pairs.

>p53(1-93) PRO to GLY middle 11
CCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAG
CGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTTCT
TCTAGTGTAGCCGTAGTTAGCCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCT
GCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGG
GTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCC
CGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGG
TATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAA
ACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGT
GGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCG
GAAGGCGAGAGTAGGGAACTGCCAGGCATCAAACTAAGCAGAAGGCCCCTGACGGATGGCCTTTTTGCGTTTCTACAAACTCTTTCTG
TGTTGTAAAACGACGGCCAGTCTTAAGCTCGGGCCCCCTGGGCGGTTCTGATAACGAGTAATCGTTAATCCGCAAATAACGTAAAAAC
CCGCTTCGGCGGGTTTTTTTTATGGGGGGAGTTTAGGGAAAGAGCATTTGTCAGAATATTTAAGGGCGCCTGTCACTTTGCTTGATATA
TGAGAATTATTTAACCTTATAAATGAGAAAAAAGCAACGCACTTTAAATAAGATACGTTGCTTTTTCGATTGATGAACACCTATAATT
AAACTATTCATCTATTATTTATGATTTTTTGTATATACAATTTCTAGTTTGTTAAAGAGAATTAAGAAAATAAATCTCGAAAATAA
TAAAGGGAAAATCAGTTTTTGATATCAAAATTATACATGTCAACGATAATACAAAATATAATACAAACTATAAGATGTTATCAGTATT
TATTATCATTTAGAATAAATTTTGTGTCGCCCTTAATTGTGAGCGGATAACAATTACGAGCTTCATGCACAGTGAAATCATGAAAAAT
TTATTTGCTTTGTGAGCGGATAACAATTATAATATGTGGAATTGTGAGCGCTCACAATTCCACAACGGTTTCCCTCTAGAAATAATTT
TGTTTAACTTTTAAGGAGGTAAAAAATGCGCGGTTCTCACCATCATCATCATCACTCTAGCGGTCTGGTCCCTCGTGGTTCTATGGAA
GAACCGCAATCGGATCCGAGCGTTGAGCCGCCACTGAGCCAGGAAACCTTCAGCGACCTGTGGAAACTGCTGCCGGAGAATAACGTTC
TGAGCGGTTTGGGTTCCCAAGCGATGGACGACCTGATGCTGAGCGGTGATGATATCGAGCAGTGGTTTACCGAAGATGGCGGCGGCGA
CGAAGCCGGCCGTATGGGTGAGGCCGCGGGCGGCGTGGCGGGTGCGGGTGCAGCGCCGACGCCGGCTGCACCGGCACCGGCTCCGAGC
TGGCCGTTGTGATAACCCCAAGGGCGACACCCCATAATTAGCCCGGGCGAAAGGCCCAGTCTTTCGACTGAGCCTTTCGTTTTATTTG
ATGCCTGGCAGTTCCCTACTCTCGCATGGGGAGTCCCCACACTACCATCGGCGCTACGGCGTTTCACTTCTGAGTTCGGCATGGGGTC
AGGTGGGACCACCGCGCTACTGCCGCCAGGCAAACAAGGGGTGTTATGAGCCATATTCAGGTATAAATGGGCTCGCGATAATGTTCAG
AATTGGTTAATTGGTTGTAACACTGACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAACCC
TGATAAATGCTTCAATAATATTGAAAAAGGAAGAATATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTT
GCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACT
GGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGC
GCGGTATTATCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACTCACCAG
TCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTT
ACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAA
CCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCGATGGCAACAACGTTGCGCAAACTATTAACTG
GCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCT
TCCGGCTGGCTGGTTTATTGCTGATAAATCCGGAGCCGGTGAGCGTGGTTCTCGCGGTATCATCGCAGCGCTGGGGCCAGATGGTAAG
CCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCAC
TGATTAAGCATTGGTAAGCGGCGCGCCATCGAATGGCGCAAAACCTTTCGCGGTATGGCATGATAGCGCCCGGAAGAGAGTCAATTCA
GGGTGGTGAATATGAAACCAGTAACGTTATACGATGTCGCAGAGTATGCCGGTGTCTCTTATCAGACCGTTTCCCGCGTGGTGAACCA
GGCCAGCCACGTTTCTGCGAAAACGCGGGAAAAAGTGGAAGCGGCGATGGCGGAGCTGAATTACATTCCCAACCGCGTGGCACAACAA
CTGGCGGGCAAACAGTCGTTGCTGATTGGCGTTGCCACCTCCAGTCTGGCCCTGCACGCGCCGTCGCAAATTGTCGCGGCGATTAAAT
CTCGCGCCGATCAACTGGGTGCCAGCGTGGTGGTGTCGATGGTAGAACGAAGCGGCGTCGAAGCCTGTAAAGCGGCGGTGCACAATCT
TCTCGCGCAACGCGTCAGTGGGCTGATCATTAACTATCCGCTGGATGACCAGGATGCCATTGCTGTGGAAGCTGCCTGCACTAATGTT
CCGGCGTTATTTCTTGATGTCTCTGACCAGACACCCATCAACAGTATTATTTTCTCCCATGAGGACGGTACGCGACTGGGCGTGGAGC
ATCTGGTCGCATTGGGTCACCAGCAAATCGCGCTGTTAGCGGGCCCATTAAGTTCTGTCTCGGCGCGTCTGCGTCTGGCTGGCTGGCA
TAAATATCTCACTCGCAATCAAATTCAGCCGATAGCGGAACGGGAAGGCGACTGGAGTGCCATGTCCGGTTTTCAACAAACCATGCAA
ATGCTGAATGAGGGCATCGTTCCCACTGCGATGCTGGTTGCCAACGATCAGATGGCGCTGGGCGCAATGCGCGCCATTACCGAGTCCG
GGCTGCGCGTTGGTGCGGATATCTCGGTAGTGGGATACGACGATACCGAAGATAGCTCATGTTATATCCCGCCGTTAACCACCATCAA
ACAGGATTTTCGCCTGCTGGGGCAAACCAGCGTGGACCGCTTGCTGCAACTCTCTCAGGGCCAGGCGGTGAAGGGCAATCAGCTGTTG
CCAGTCTCACTGGTGAAAAGAAAAACCACCCTGGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAATGCAGC
TGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAGTGACTCATGACCAAAATCCCTTAACGTGAGTTACGCGCGCGTCGTTCCACTGA
GCGTCAGAC

FIG. 7. **FASTA sequence of pJ404 plasmid containing M11p53.** Sequences are highlighted according to the following legend: *Ori_puC*, *term_rpoc*, *Term_bla*, *LacO1/P_T5_Inducible*, *P_T5_inducible*, *lacO/P_T5_inducible*, *5'UTR_T7_10A*, *RBS*, *HIS*, *AMP_p*, *Ampicilin-r*, *P_lacI*, *lacI*. This plasmid contains 4321 total base pairs.

>p53(1-93) PRO to GLY last 11
CCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAG
CGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTTCT
TCTAGTGTAGCCGTAGTTAGCCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCT
GCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGG
GTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCC
CGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGG
TATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAA
ACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGT
GGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCG
GAAGGCGAGAGTAGGGAACTGCCAGGCATCAAACTAAGCAGAAGGCCCCTGACGGATGGCCTTTTTGCGTTTCTACAAACTCTTTCTG
TGTTGTAAAACGACGGCCAGTCTTAAGCTCGGGCCCCCTGGGCGGTTCTGATAACGAGTAATCGTTAATCCGCAAATAACGTAAAAAC
CCGCTTCGGCGGGTTTTTTTTATGGGGGGAGTTTAGGGAAAGAGCATTTGTCAGAATATTTAAGGGCGCCTGTCACTTTGCTTGATATA
TGAGAATTATTTAACCTTATAAATGAGAAAAAAGCAACGCACTTTAAATAAGATACGTTGCTTTTTCGATTGATGAACACCTATAATT
AAACTATTCATCTATTATTTATGATTTTTTGTATATACAATTTCTAGTTTGTTAAAGAGAATTAAGAAAATAAATCTCGAAAATAA
TAAAGGGAAAATCAGTTTTTGATATCAAAATTATACATGTCAACGATAATACAAAATATAATACAAACTATAAGATGTTATCAGTATT
TATTATCATTTAGAATAAATTTTGTGTCGCCCTTAATTGTGAGCGGATAACAATTACGAGCTTCATGCACAGTGAAATCATGAAAAAT
TTATTTGCTTTGTGAGCGGATAACAATTATAATATGTGGAATTGTGAGCGCTCACAATTCCACAACGGTTTCCCTCTAGAAATAATTT
TGTTTAACTTTTAAGGAGGTAAAAAATGCGCGGTTCTCATCATCACCATCATCACAGCAGCGGTCTGGTCCCTCGTGGCAGCTTGGGC
TGGAGCGGTGCAGGCGCGGGCGCGGCTGGTACCGGTGCGGCGGGTGCGGGTGCCGTTGGTGGCGCAGCTGAGGGCATGCGTCCGGCCG
AGGATCCGGGTCCGGACGAAACGTTTTGGCAAGAAATCGACGATCCGAGCCTGATGCTGGACGACATGGCACAGTCGCCGCTGCCGAG
CCTGGTGAACAATGAACCGTTGCTGAAATGGCTGGATAGCTTCACCGAGCAGTCCCTGCCGCCAGAAGTGAGCCCGGATTCCCAACCG
GAAGAAATGTGATAACCCCAAGGGCGACACCCCATAATTAGCCCGGGCGAAAGGCCCAGTCTTTCGACTGAGCCTTTCGTTTTATTTG
ATGCCTGGCAGTTCCCTACTCTCGCATGGGGAGTCCCCACACTACCATCGGCGCTACGGCGTTTCACTTCTGAGTTCGGCATGGGGTC
AGGTGGGACCACCGCGCTACTGCCGCCAGGCAAACAAGGGGTGTTATGAGCCATATTCAGGTATAAATGGGCTCGCGATAATGTTCAG
AATTGGTTAATTGGTTGTAACACTGACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAACCC
TGATAAATGCTTCAATAATATTGAAAAAGGAAGAATATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTT
GCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACT
GGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGC
GCGGTATTATCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACTCACCAG
TCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTT
ACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAA
CCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCGATGGCAACAACGTTGCGCAAACTATTAACTG
GCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCT
TCCGGCTGGCTGGTTTATTGCTGATAAATCCGGAGCCGGTGAGCGTGGTTCTCGCGGTATCATCGCAGCGCTGGGGCCAGATGGTAAG
CCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCAC
TGATTAAGCATTGGTAAGCGGCGCGCCATCGAATGGCGCAAAACCTTTCGCGGTATGGCATGATAGCGCCCGGAAGAGAGTCAATTCA
GGGTGGTGAATATGAAACCAGTAACGTTATACGATGTCGCAGAGTATGCCGGTGTCTCTTATCAGACCGTTTCCCGCGTGGTGAACCA
GGCCAGCCACGTTTCTGCGAAAACGCGGGAAAAAGTGGAAGCGGCGATGGCGGAGCTGAATTACATTCCCAACCGCGTGGCACAACAA
CTGGCGGGCAAACAGTCGTTGCTGATTGGCGTTGCCACCTCCAGTCTGGCCCTGCACGCGCCGTCGCAAATTGTCGCGGCGATTAAAT
CTCGCGCCGATCAACTGGGTGCCAGCGTGGTGGTGTCGATGGTAGAACGAAGCGGCGTCGAAGCCTGTAAAGCGGCGGTGCACAATCT
TCTCGCGCAACGCGTCAGTGGGCTGATCATTAACTATCCGCTGGATGACCAGGATGCCATTGCTGTGGAAGCTGCCTGCACTAATGTT
CCGGCGTTATTTCTTGATGTCTCTGACCAGACACCCATCAACAGTATTATTTTCTCCCATGAGGACGGTACGCGACTGGGCGTGGAGC
ATCTGGTCGCATTGGGTCACCAGCAAATCGCGCTGTTAGCGGGCCCATTAAGTTCTGTCTCGGCGCGTCTGCGTCTGGCTGGCTGGCA
TAAATATCTCACTCGCAATCAAATTCAGCCGATAGCGGAACGGGAAGGCGACTGGAGTGCCATGTCCGGTTTTCAACAAACCATGCAA
ATGCTGAATGAGGGCATCGTTCCCACTGCGATGCTGGTTGCCAACGATCAGATGGCGCTGGGCGCAATGCGCGCCATTACCGAGTCCG
GGCTGCGCGTTGGTGCGGATATCTCGGTAGTGGGATACGACGATACCGAAGATAGCTCATGTTATATCCCGCCGTTAACCACCATCAA
ACAGGATTTTCGCCTGCTGGGGCAAACCAGCGTGGACCGCTTGCTGCAACTCTCTCAGGGCCAGGCGGTGAAGGGCAATCAGCTGTTG
CCAGTCTCACTGGTGAAAAGAAAAACCACCCTGGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAATGCAGC
TGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAGTGACTCATGACCAAAATCCCTTAACGTGAGTTACGCGCGCGTCGTTCCACTGA
GCGTCAGAC

FIG. 8. **FASTA sequence of pJ404 plasmid containing C11p53.** Sequences are highlighted according to the following legend: *Ori_puC*, *term_rpoc*, *Term_bla*, *LacO1/P_T5_Inducible*, *P_T5_inducible*, *lacO/P_T5_inducible*, *5'UTR_T7_10A*, *RBS*, *HIS*, *AMP_p*, *Ampicilin-r*, *P_lacI*, *lacI*. This plasmid contains 4321 total base pairs.

1 MRGSHHHHHH SSGLVPRGSM EEPQSDPSVE PPLSQETFSD LWKLLPENNV

60 LSPLPSQAMD DLMLSPDDIE QWFTEDPGPD EAPRMPEAAP PVAPAPAAPT

110 PAAPAPAPSW PL

FIG. 9. **Amino acid sequence of p53(1-93).** The portion in red indicates the histidine affinity tag used to purify and be subsequently cleaved by thrombin. The portion in blue is the remainder after thrombin cleavage.

1 MRGSHHHHHH SSGLVPRGSL PWSPAPAPAA PTPAAPAPAV PPAAEPMRPA

60 EDPGPDETFW QEIDDPSLML DDMAQSPLPS LVNNEPLLKW LDSFTEQSLP

110 PEVSPDSQPE EM

FIG. 10**. Amino acid sequence of p53(93-1).** The portion in red indicates the histidine affinity tag used to purify and be subsequently cleaved by thrombin. The portion in blue is the remainder after thrombin cleavage.

1 MRGSHHHHHH SSGLVPRGSL PWSPAPAPAA PTPAAPAPAV PPAAEPMRGA

60 EDGGGDETFW QEIDDGSLML DDMAQSGLGS LVNNEGLLKW LDSFTEQSLG

110 GEVSGDSQGE EM

FIG. 11. **Amino acid sequence of RF11p53.** The portion in red indicates the histidine affinity tag used to purify and be subsequently cleaved by thrombin. The portion in blue is the remainder after thrombin cleavage. The orange glycines indicate PRO → GLY substitution sites.

1 MRGSHHHHHH SSGLVPRGSL PWSPAPAPAA PTPAAGAGAV GGAAEGMRGA

60 EDGGGDETFW QEIDDGSLML DDMAQSGLGS LVNNEPLLKW LDSFTEQSLP

110 PEVSPDSQPE EM

FIG. 12. **Amino acid sequence of RM11p53.** The portion in red indicates the histidine affinity tag used to purify and be subsequently cleaved by thrombin. The portion in blue is the remainder after thrombin cleavage. The orange glycines indicate PRO → GLY substitution sites.

1 MRGSHHHHHH SSGLVPRGSL GWSGAGAGAA GTGAAGAGAV GGAAEGMRPA

60 EDPGPDETFW QEIDDPSLML DDMAQSPLPS LVNNEPLLKW LDSFTEQSLP

110 PEVSPDSQPE EM

FIG. 13. **Amino acid sequence of RL11p53.** The portion in red indicates the histidine affinity tag used to purify and be subsequently cleaved by thrombin. The portion in blue is the remainder after thrombin cleavage. The orange glycines indicate PRO → GLY substitution sites.

1 MRGSHHHHHH SSGLVPRGSM EEGQSDGSVE GGLSQETFSD LWKLLGENNV

60 LSGLGSQAMD DLMLSGDDIE QWFTEDGGGD EAGRMPEAAP PVAPAPAAPT

110 PAAPAPAPSW PL

FIG. 14. **Amino acid sequence of N11p53.** The portion in red indicates the histidine affinity tag used to purify and be subsequently cleaved by thrombin. The portion in blue is the remainder after thrombin cleavage. The orange glycines indicate PRO → GLY substitution sites.

1 MRGSHHHHHH SSGLVPRGSM EEPQSDPSVE PPLSQETFSD LWKLLPENNV
60 LSGLGSQAMD DLMLSGDDIE QWFTEDGGGD EAGRMGEAAG GVAGAGAAPT 110
PAAPAPAPSW PL

FIG. 15. **Amino acid sequence of M11p53.** The portion in red indicates the histidine affinity tag used to purify and be subsequently cleaved by thrombin. The portion in blue is the remainder after thrombin cleavage. The orange glycines indicate PRO → GLY substitution sites.

1 MRGSHHHHHH SSGLVPRGSM EEPQSDPSVE PPLSQETFSD LWKLLPENNV
60 LSPLPSQAMD DLMLSPDDIE QWFTEDPGPD EAPRMGEAAG GVAGAGAAGT
110 GAAGAGAGSW GL

FIG. 16. **Amino acid sequence of C11p53.** The portion in red indicates the histidine affinity tag used to purify and be subsequently cleaved by thrombin. The portion in blue is the remainder after thrombin cleavage. The orange glycines indicate PRO → GLY substitution sites.

Materials

All water for experiments was filtered and deionized using a EMD Millipore Milli-Q Integral 3 water purifier (Billerica, MA). Glassware and other equipment was sterilized using a Hirayama HICLAVE HV-50 autoclave (Kasukabe-Shi, Japan). All agar plates were obtained from Teknova (Hollister, CA). Sonication was performed using a Branson Sonifer S-450A (Danbury, CT). Incubation of bacteria cultures was performed using a Fisher Scientific MaxQ 500 floor-model shaker (Waltham, MA). Column chromatography was performed using a Bio-Rad Biologic LP chromatography system (Hercules, CA). The pH electrodes used were either the Beckman Coulter Theta 510 pH meter (Brea, CA) or a Thermo Scientific Orion Star A111 pH meter (Waltham, MA). Reagents were weighed using a A&D GH-200 analytical balance (Tokyo, Japan) or a PA84C Pioneer from Ohaus (Parsippany, NJ). A Welch DryFast 2032 Ultra Diaphragm Pump was used to degass column media and for vacuum filtration (Niles, IL). CD spectroscopy was performed using a Jasco J-710 spectropolarimeter with a Jasco PFD-425S Peltier (Easton, MD).

Transformation

The genes containing the sequences above were cloned into the pJ404 expression vector for the N11p53, M11p53, and C11p53 proteins and into the pJ414 expression vector for p53(1-93), p53(93-1), RF11p53, RM11p53, RL11p53, from DNA 2.0 (Menlo Park, CA). Glycerol stocks created by Dr. Steve Whitten were used for wild-type expression. The pJ404 expression vector contains a gene that confers ampicillin resistance as well as an IPTG-

inducible T5 promoter. The pJ414 expression vector contains an ampicillin resistance gene and an IPTG inducible T7 promoter. The plasmids were aliquoted and diluted with autoclaved, deionized water to a concentration of 20 ng/μL and frozen at -80 °C. The plasmids were then combined with competent BL21(DE3) *E. coli* cells by mixing 10 ng of DNA with 100 μL of bacteria.

Transformation was completed by allowing the mixture to incubate on ice for 30 minutes, subjecting to heat-shock at 42 °C for 2 minutes, and incubating on ice for 2 minutes. An outgrowth was prepared by adding 1 mL of either SOC or 2xYT broth, and allowed to incubate at 37 °C with 225 RPM shaking for 60 minutes. Cells were plated on LB agar containing ampicillin and LB agar without ampicillin as a control by taking 50 mL of outgrowth and spreading on the plate, and allowed to incubate overnight at 37 °C and shaking at 225 RPM.

Expression

A single colony was obtained from previously transformed plates and allowed to grow in 3 mL of 2xYT broth containing 100 μg/mL of ampicillin overnight at 37 °C. The following day, 1 mL of overnight growth culture was added to 1L of 2xYT media containing 100 μg/mL of ampicillin and incubated at 37 °C and 225 RPM until reaching an optical density ~ 0.6 - 0.7 at 600 nm to express during the log phase of bacterial growth. After reaching the log phase of growth, 0.5 mM IPTG was added to induce expression for 4 hours. The culture was centrifuged at 4 °C, 30,240 x g for 15 minutes in a F14-6x250y rotor. The supernatant was decanted and the precipitate cell pellets frozen at -80 °C.

Purification

Cell pellets were removed from the freezer and solubilized in 6 M GuHCl, 10 mM Tris-HCl, 100 mM Na_2PO_4, pH 8.0. The pellets were lysed through four cycles of sonication for two minutes each, and centrifuged at 4 °C, 30,240 x g for 1 hour. The samples were loaded onto a His-Select Nickel Affinity gel column from Sigma-Aldrich (St. Louis, MO) equilibrated in 6 M GuHCl, 10 mM Tris-HCl, 100 mM Na_2PO_4, pH 8.0. Column washes were performed using 1) 80 mL of 6 M GuHCl, 10 mM Tris-HCl, 100 mM Na_2PO_4, pH 8.0; 2) 80 mL 10 mM Tris-HCl, 100 mM Na_2PO_4, pH 8.0 3) 80 mL 10 mM Tris-HCl, 100 mM Na_2PO_4, 10 mM imidazole, pH 8.0; and eluted with 60 mL 10 mM Tris-HCl, 100 mM Na_2PO_4, 350 mM imidazole, pH 4.3. The eluent was dialyzed overnight in 4L 20 mM Tris-HCl, 100 mM NaCl, pH 8.

Following overnight dialysis, the eluent was treated with 10 units of thrombin from Sigma-Aldrich (St. Louis, MO) for either 4 hours (plasmids containing T5 promoter) or 16 hours (plasmids containing T7 promoter) to digest the poly-histidine tag. The samples were loaded onto degassed DEAE gel column from GE Healthcare (Little Chalfont, UK) equilibrated in 20 mM NaAc, 25 mM NaCl, pH 4.8. Column washes were performed using 1) 60 mL 20 mM NaAc, 25 mM NaCl, pH 4.8; 2) 60 mL 20 mM NaAc, 50 mM NaCl, pH 4.8; 3) 60 mL 20 mM NaAc, 150 mM NaCl, pH 4.8; and eluted with 60 mL 20 mM NaAc, 400 mM NaCl, pH 4.8. The eluent was dialyzed overnight in 4L 10 mM Na_2HPO_4, 100 mM NaCl, pH 7.

Gel Electrophoresis

Gel electrophoresis was performed to verify complete digestion of the histidine tag and show purity of product. Gels were 18 Well 4-20% Tris-HCl, 1.0 mm Criterion Precast Gels from Bio-Rad (Hercules, CA). Samples for loading were mixed with equal volumes of 19:1 ratio of 2x Laemmli Sample Buffer and β-mercaptoethanol. Sample/buffer mix of 20 µL was loaded into the gel buffered with TGS. Unused lanes were loaded with 10 µL of buffer mix. A ladder of 4 µL Precision Plus Protein^{TM} All Blue Standards from Bio-Rad (Hercules, CA) was loaded for comparison to known standards. A voltage of 100 V was applied for 10 minutes, then voltage was increased to 200 V until the migration was complete. Gels were stained with 10% Acetic Acid, 40% Methanol, 0.05% Coomassie R-250 for approximately 2 hours, then destained with 10% Acetic Acid, 40% Methanol until bands began to appear, approximately 15 minutes. Gels were imaged using either a ChemiDoc^{TM} XRS+ System from Bio-Rad (Hercules, CA). To estimate purity, gels were analyzed using the software ImageJ [45]. Images were first converted to 32-bit greyscale, then individual lanes were selected to generate a plot profile. The area underneath the image intensity plot profile is a semi-quantitative way to determine gel purity.

Circular Dichroism Spectroscopy

Circular Dichroism was performed using a J-710 Spectropolarimeter equipped with a PFT-425S Peltier from Jasco (Oklahoma City, OK). Samples were diluted to a concentration indicated in figure legends and a total volume of 300 µL into a J/0556 Cuvette from Jasco (Oklahoma City, OK). Nitrogen gas was used to purge the spectropolarimeter, applied at a

pressure of ~20 PSI. A total of eight wavelength scans between 195-245 nm were measured for 10° increments between 5 °C and 85 °C, and subtracted from a baseline measurement of 10 mM Na_2HPO_4, 100 mM NaCl, pH 7.

Size Exclusion Chromatography

To determine R_h, samples were measured through size exclusion chromatography using Sephadex G-75 Media from GE Healthcare (Little Chalfont, UK) hydrated in degassed 10 mM Na_2HPO_4, 100 mM NaCl, pH 7. The elution time of p53(1-93) samples were compared to the elution times of Blue Dextran from Sigma (St. Louis, MO) and DNP-Aspartate from Research Organics (Cleveland, Ohio). The relative elution times for each sample were calculated into K_d ratios;

$$K_D = \frac{V_e - V_0}{V_t - V_0}$$

Where V_e represents the elution time of the sample, V_0 represents the elution time of Blue Dextran, and V_t represents the elution time of DNP-Aspartate. This was calculated for each of the p53 variants, as well as horse heart myoglobin (Sigma-Aldrich M1882), Bovine Carbonic Acid (Sigma-Aldrich C5023), *Staphylococcus* nuclease (Previously purified by Lance English, according to Whitten and Garcia-moreno 2000), and Chicken Egg Albumen (Thermo-Fisher Scientific Acros 40045), hydrated in 10 mM Na_2HPO_4, 100 mM NaCl, pH 7, which were run as standards. The R_h for the SEC protein standards were determined by measuring the distance between the two furthest α-carbons of a single subunit in the protein

crystal structure PDB file. This method has been shown to be a reasonable approximation for prediction of R_h of folded proteins and for computationally generated IDP structures [25, 29]. The R_h of protein standards are shown in Table 1.

Table 1. **Calculated protein crystallographic R_h.** Values were derived from PDB files, and calculating the largest distance between two alpha carbons located in a single subunit.

PDB	NAME	RESIDUES	R_H
1OVA	Chicken Egg Albumen	386	35.92
1V9E	Bovine Carbonic Anyhydrase	259	27.40
1STN	*Staphylococcal* Nuclease	149	21.20
2O58	Horse Heart Myoglobin	153	21.83
4R5S	Bovine Serum Albumin	583	42.07

III. RESULTS AND DISCUSSION

Nickel Affinity Chromatography

Proteins were successfully purified through nickel affinity chromatography, as indicated by the clear peak after elution buffers in the chromatograms, which indicates an absorbance of 280 nm, indicative of presence of a protein. An example of a nickel affinity chromatograph for the purification of p53(93-1) is displayed in **Figure 17**.

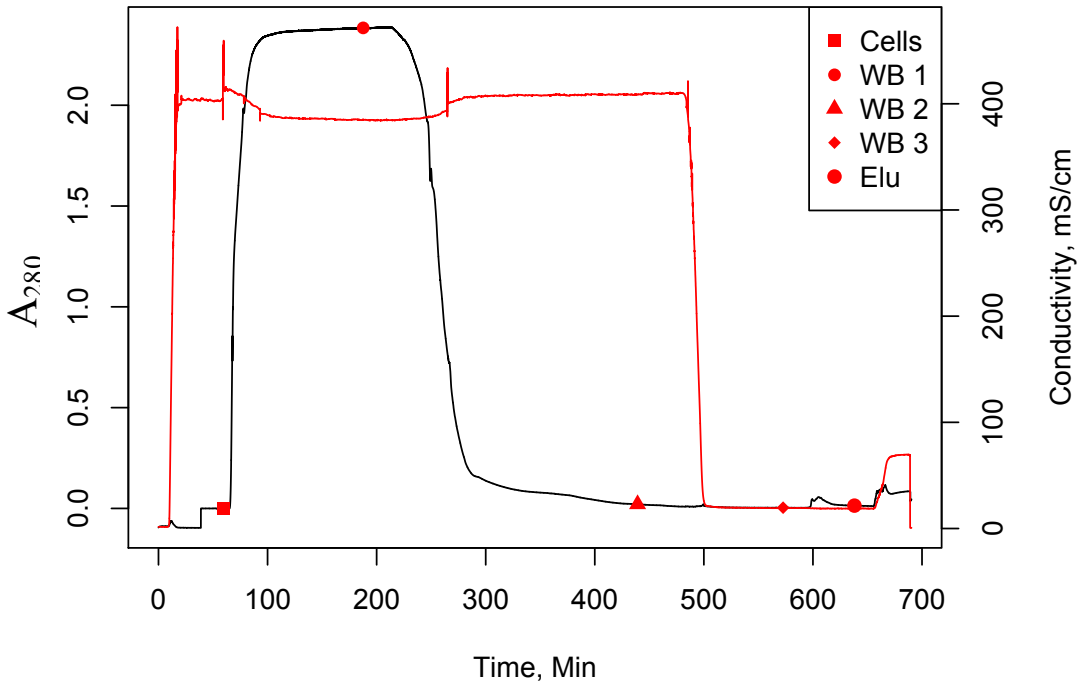


FIG. 17. **Nickel affinity column chromatogram and conductivity readings for p53(93-1).** The black line represents absorbance at 280 nm. The red line represents conductivity measurmeents. Buffer markings; *Cells*, clarified cell lysate from expression, T = 59.23 min; *WB 1,* wash buffer containing: 6 M GuHCl, 10 mM Tris-HCl, 100 mM Na_2PO_4, pH 8.0, T = 187.66 min; *WB 2,* wash buffer containing: 10 mM Tris-HCl, 100 mM Na_2PO_4, pH 8.0, T = 439.23 min; *WB 3,* wash buffer containing: 10 mM Tris-HCl, 100 mM Na_2PO_4, 10 mM imidazole, pH 8.0, T = 572.64 min; *Elu*, elution buffer containing: 20 mM NaAc, 400 mM NaCl, pH 4.8, T = 638.05 min. The peak after the addition of the elution buffer is the target protein.

Anion Exchange Chromatography

Following thrombin digestion of the poly-histidine tag and verifying complete digestion with a gel electrophoresis experiment, the samples were successfully run through a DEAE column, indicating a clear elution peak for each protein by measuring the absorption of light at 280 nm. An example of an anion exchange chromatograph for the purification of p53(93-1) is displayed in **Figure 18**.

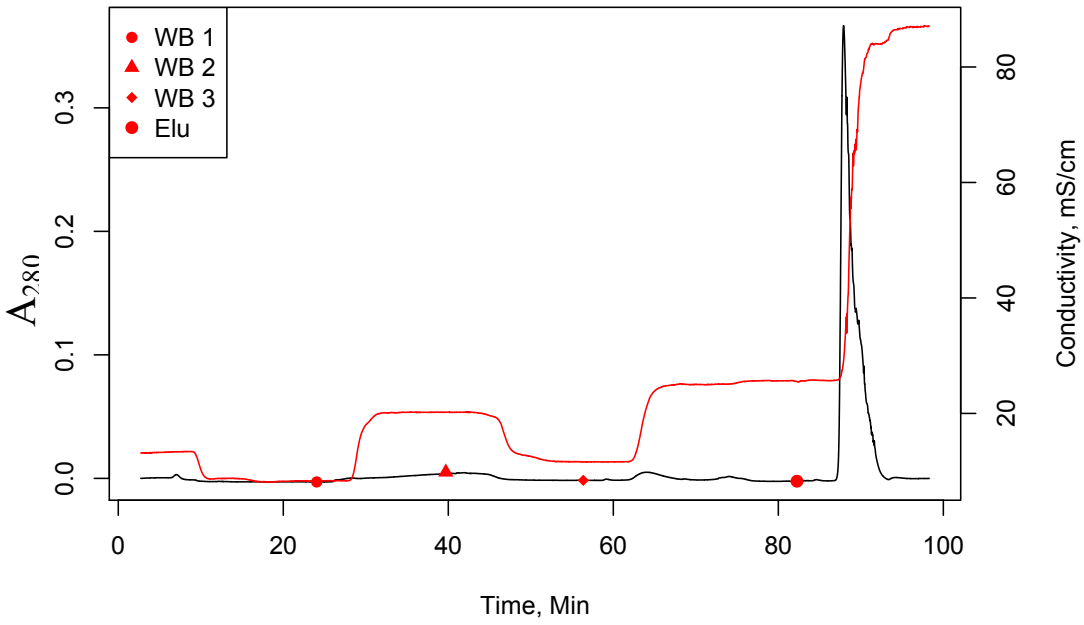


FIG. 18. **DEAE chromatogram and conductivity readings for p53(93-1).** The black line represents absorbance at 280 nm. The red line represents conductivity measurements. Buffer markings; *WB 1,* wash buffer containing: 20 mM NaAc, 25 mM NaCl, pH 4.8, T = 24.05 min; *WB 2,* wash buffer containing: 20 mM NaAc, 50 mM NaCl, pH 4.8, T = 39.7 min; *WB 3,* wash buffer containing: 20 mM NaAc, 150 mM NaCl, pH 4.8, T = 56.36 min; *Elu*, elution buffer containing: 20 mM NaAc, 400 mM NaCl, pH 4.8, T = 82.26 min. The peak after the addition of the elution buffer is the target protein.

Gel Electrophoresis

The wild type p53(1-93), N11p53, M11p53, and C11p53 were purified and analyzed on a gel stained with Coomassie as shown in **Figure 19**. The retro variants, p53(93-1), RF11p53, RM11p53, and RL11p53 were also purified and run on a gel, as seen in **Figure 20**. Gel electrophoresis migration depends on electrophoretic mobility and hydrodynamic size, but protein standard ladders are indicated by apparent molecular weight. As expected, due to the higher molecular mass, the wild-type migrated the least, appearing around the standard indicated as 20 kDa. The C11p53 variant migrated the furthest, followed by M11p53, and N11p53 migrated the least out of the proline → glycine variants. The retro variants behaved similarly to their native isomers; the least migrated, as expected, was the retrop53(93-1), with the other retro variants following the same order of migration as their native isomers; RL11p53 migrated the furthest, followed by RM11p53, then RF11p53.

Purity analysis was performed on all variants. The purity analysis of gel electrophoresis of RM11 is shown in **Figure 21**. Gel electrophoresis results appear to display distinct hierarchy of electrophoretic mobility, where p53(1-93) and its retro variant migrate similarly just above an apparently molecular mass of 20 kDa, N11p53/RF11p53 migrate above M11p53/RM11p53 which migrates above C11p53/RL11p53. The purity analysis indicates that the proteins are pure at a level of 95%.

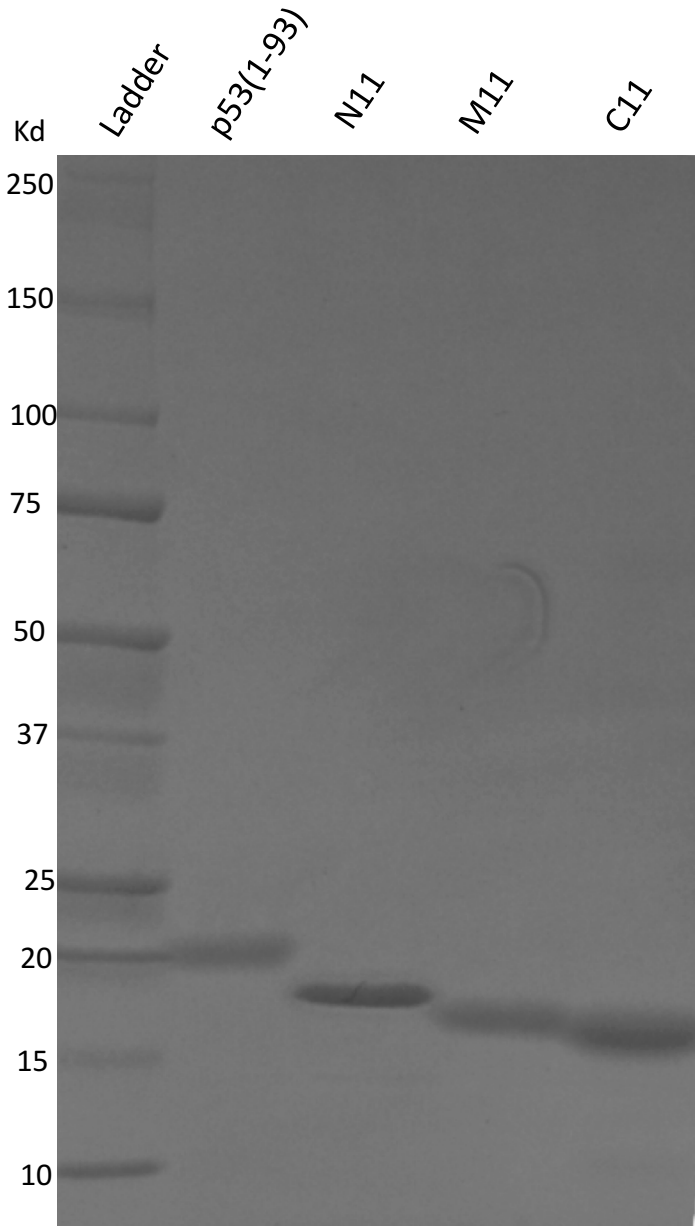


FIG. 19. **SDS-PAGE gel of purified p53(1-93), N11p53, M11p53, and C11p53.** Run on a 4-20% Tris-HCl Criterion Precast Gel with TGS, a tris-glycine running buffer. From left to right; *Lane 1*, Bio-Rad Precision Plus Protein Standards. *Lane 2,* wild-type p53. *Lane 3,* N11 p53. *Lane 4,* M11 p53. *Lane 5,* C11 p53.

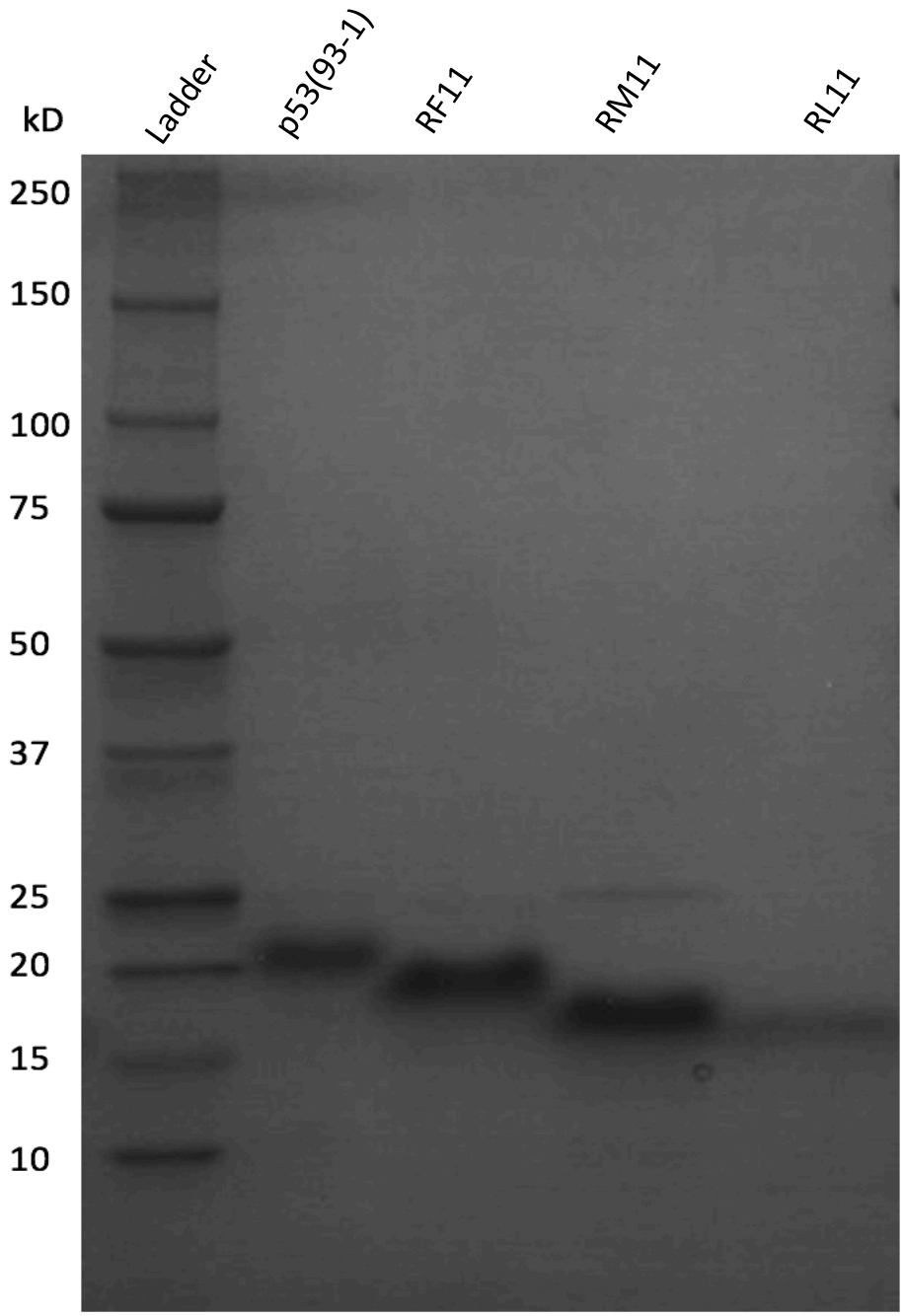


FIG. 20. **SDS-PAGE gel of purified p53(93-1), RF11p53, RM11p53, and RL11p53.** Run on a 4-20% Tris-HCl Criterion Precast Gel with TGS, a tris-glycine running buffer. From left to right; *Lane 1*, Bio-Rad Precision Plus Protein Standards. *Lane 2,* retro p53(93-1). *Lane 3,* RF11p53. *Lane 4,* RM11p53. *Lane 5,* RL11 p53.

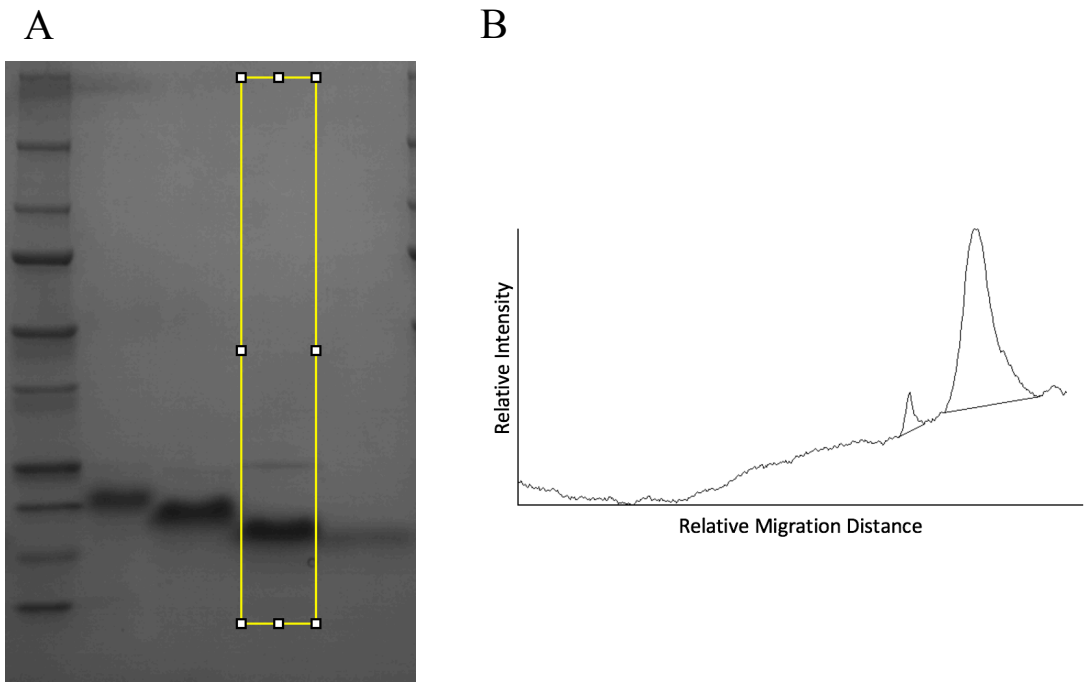
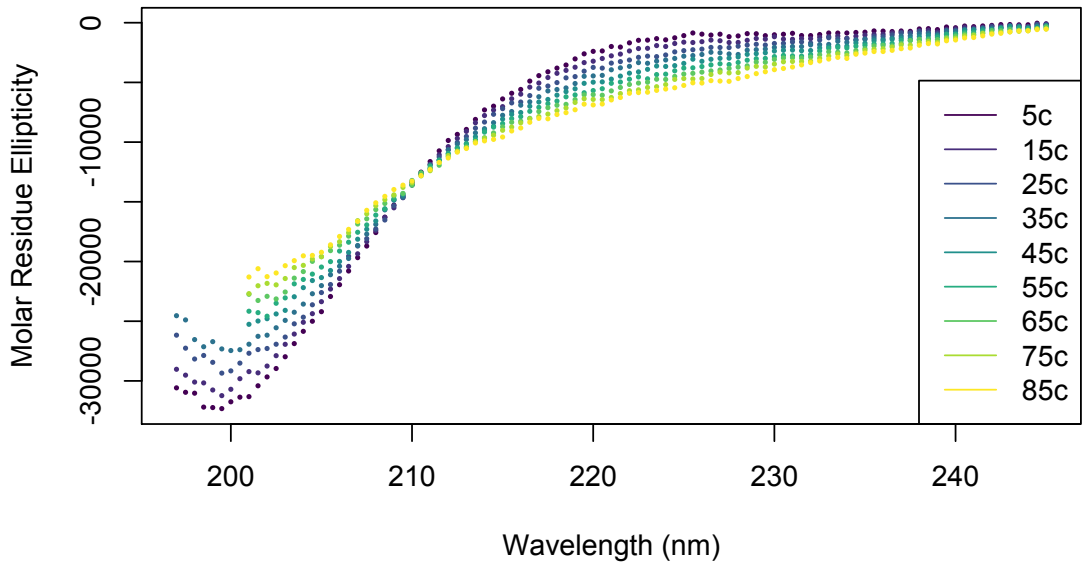


FIG. 21. **Protein purity calculations on RM11 using ImageJ software.** The boxed region in yellow indicates the region where the plot profile is measured. The lighter band at an apparent migration of around 25 kDa and the darker band at an apparent migration of around 17 kDa are represented by the smaller and larger peak, respectively. The peak area is proportional to the amount of protein present. The smaller peak has a relative area of 509.284 and the larger peak has a relative area of 8804.267, indicating a relative purity of 95%.

Circular Dichroism Spectroscopy

Each of the peptides were analyzed for secondary structure contribution using circular dichroism spectroscopy, as shown in **Figures 22-30**. As described by Schaub et. al, the peak at 220 nm indicates the presence of the PP_{II} structure [42]. For all spectra, the linear correlation of $R^2 > 0.98$ indicates the noncooperativity of thermal unfolding, as observed previously by Schaub et. al. The peak was the highest with the wild-type, as suspected due to higher PP_{II} propensity when calculated with the scale according to Elam et. al [28]. However, the N11p53 variant appears to display greater PP_{II} structure then C11p53 and M11p53. Surprisingly, all native direction variants had significantly greater PP_{II} structure than each of the retro counterparts, as indicated by the peaks at 220 nm. As predicted by all PP_{II} scales, the retro p53 had more PP_{II} structure than the other retro variants, but this is far less pronounced than the difference between the native direction wild type and variants. The RF11, RM11, and RL11 variants all displayed significantly diminished PP_{II} structure relative to their natively oriented isomers.



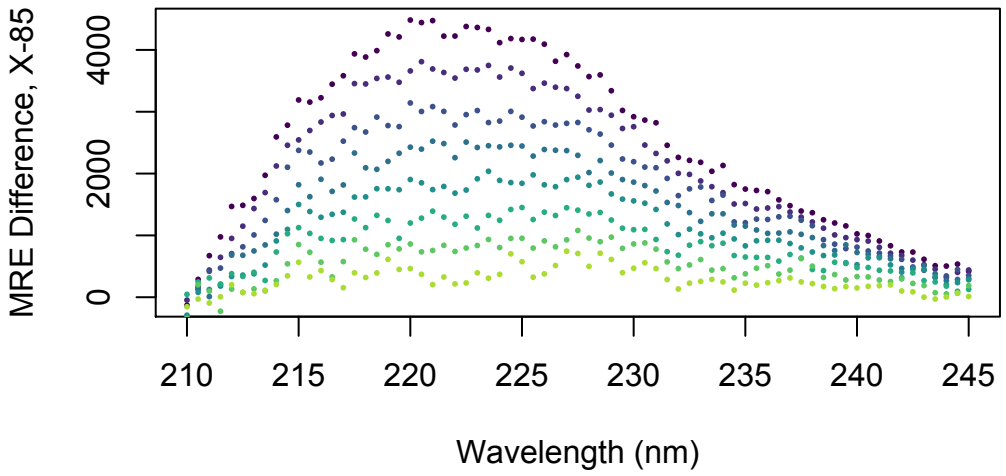


FIG. 22. **Temperature dependent CD spectra for p53(1-93).** Buffer consisted of 10mM sodium phosphate and 100mM sodium chloride at pH 7. *Top;* Molar residual ellipticity (MRE) is measured in units of deg cm^2 $dmol^{-1}$ res^{-1}. Concentration of 16 µM for this peptide. Inset: Average value of MRE from 220-222 nm for all trials, subtracted from the measurement at 85°C. *Bottom;* The difference in MRE for each wavelength between 210 and 240 nm from 85°C spectrum.

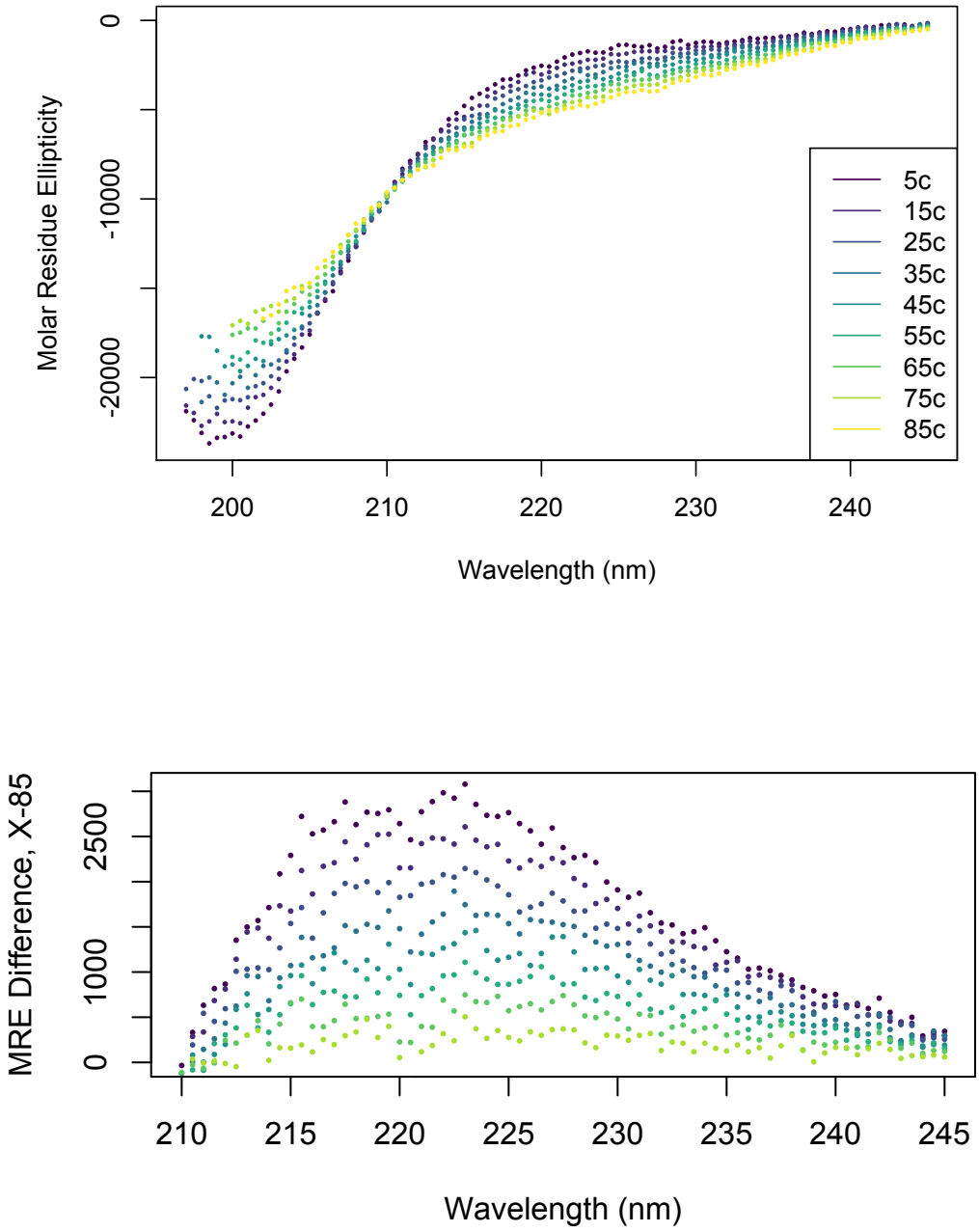


FIG. 23. **Temperature dependent CD spectra for N11p53(1-93).** Buffer consisted of 10mM sodium phosphate and 100mM sodium chloride at pH 7. *Top;* Molar residual ellipticity (MRE) is measured in units of deg cm^2 $dmol^{-1}$ res^{-1}. Concentration of 18 μM for this peptide. *Bottom;* The difference in MRE for each wavelength between 210 and 240 nm from 85°C spectrum.

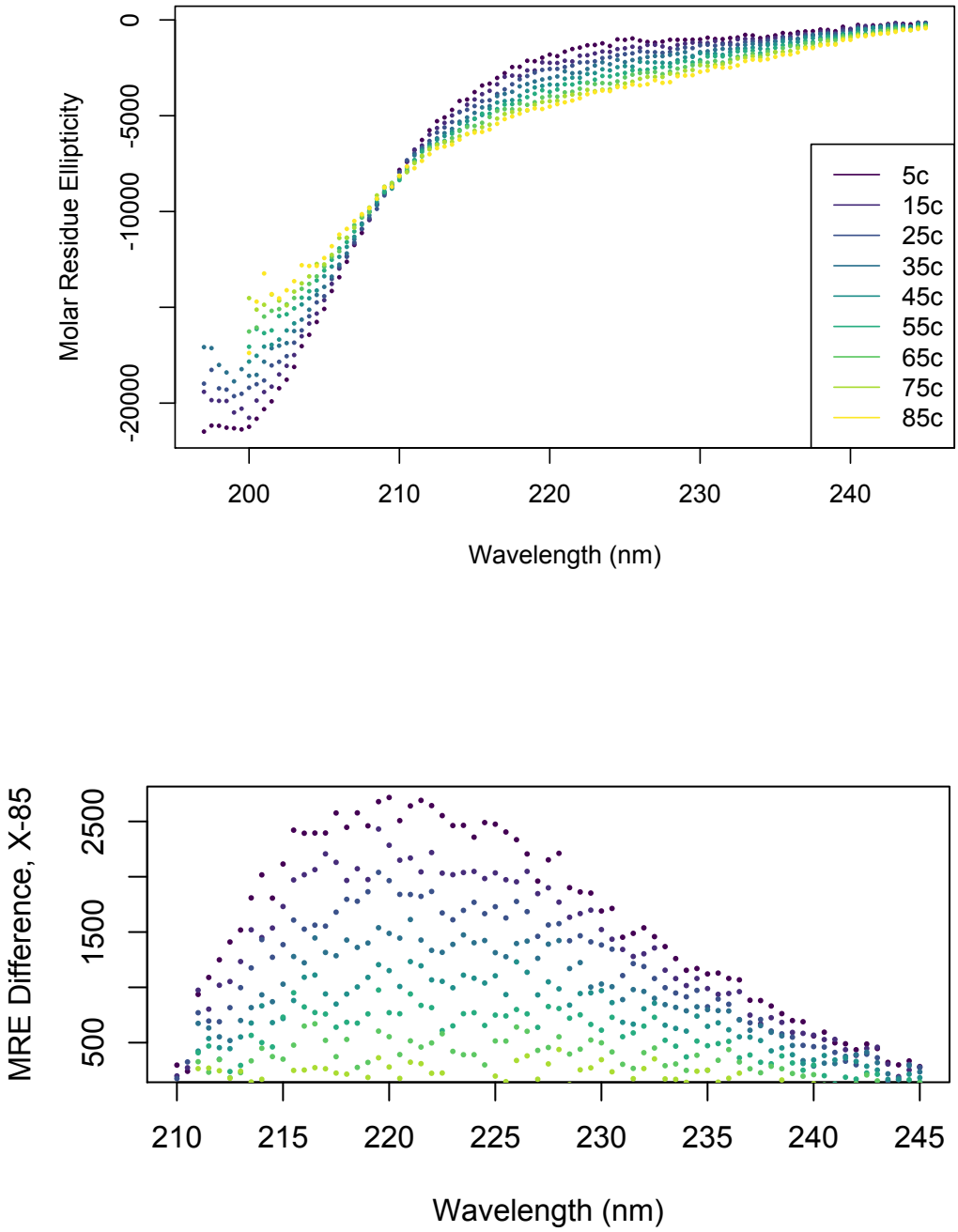


FIG. 24. **Temperature dependent CD spectra for M11p53(1-93).** Buffer consisted of 10mM sodium phosphate and 100mM sodium chloride at pH 7. *Top;* Molar residual ellipticity (MRE) is measured in units of deg cm^2 $dmol^{-1}$ res^{-1}. Concentration of 18 μM for this peptide. *Bottom;* The difference in MRE for each wavelength between 210 and 245 nm from 85°C spectrum, to show melting of PP_{II} helix.

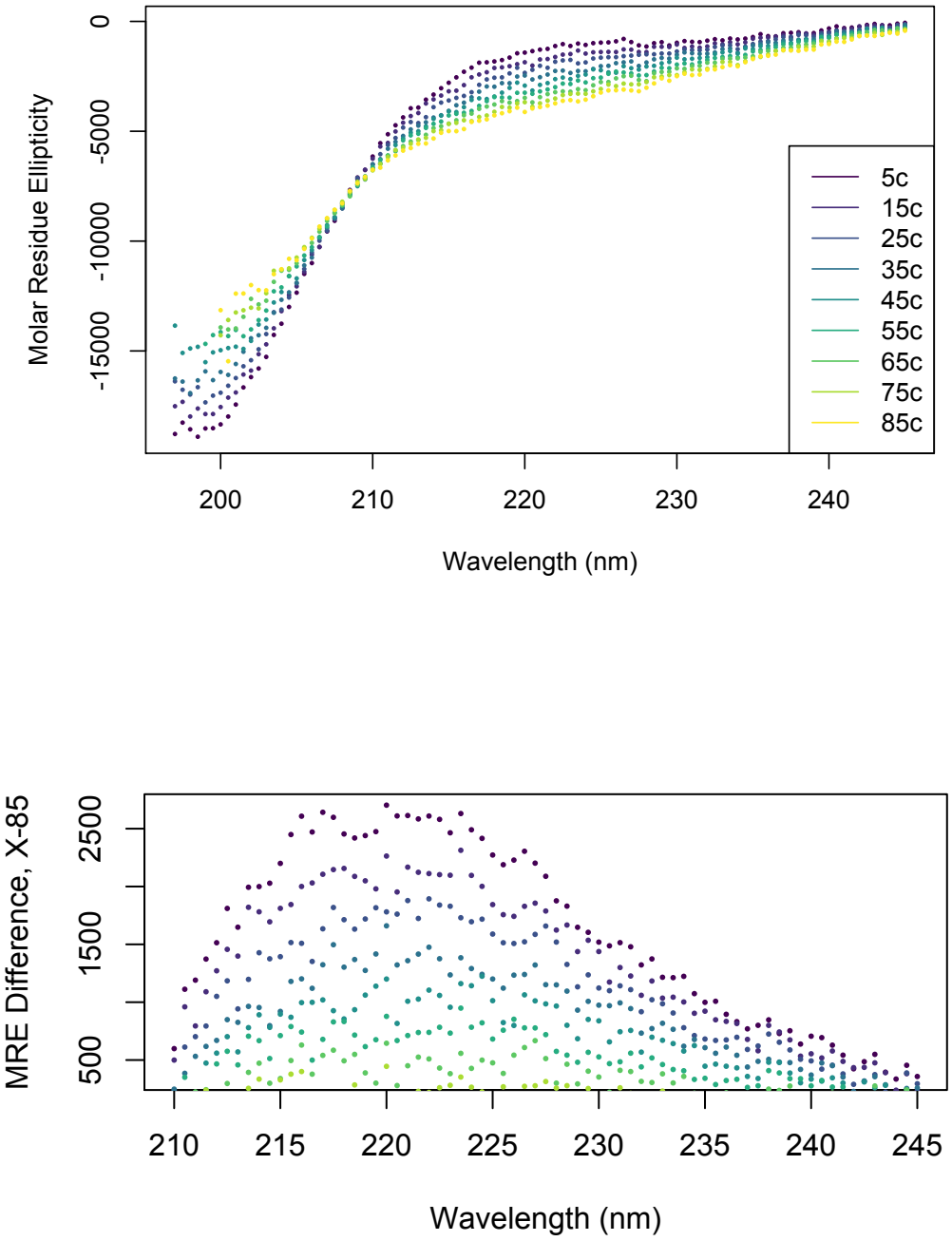


FIG. 25. **Temperature dependent CD spectra for C11p53(1-93).** Buffer consisted of 10mM sodium phosphate and 100mM sodium chloride at pH 7. *Top;* Molar residual ellipticity (MRE) is measured in units of deg cm^2 $dmol^{-1}$ res^{-1}. Concentration of 18 μM for this peptide. *Bottom;* The difference in MRE for each wavelength between 210 and 240 nm from 85°C spectrum, to show melting of PP_{II} helix.

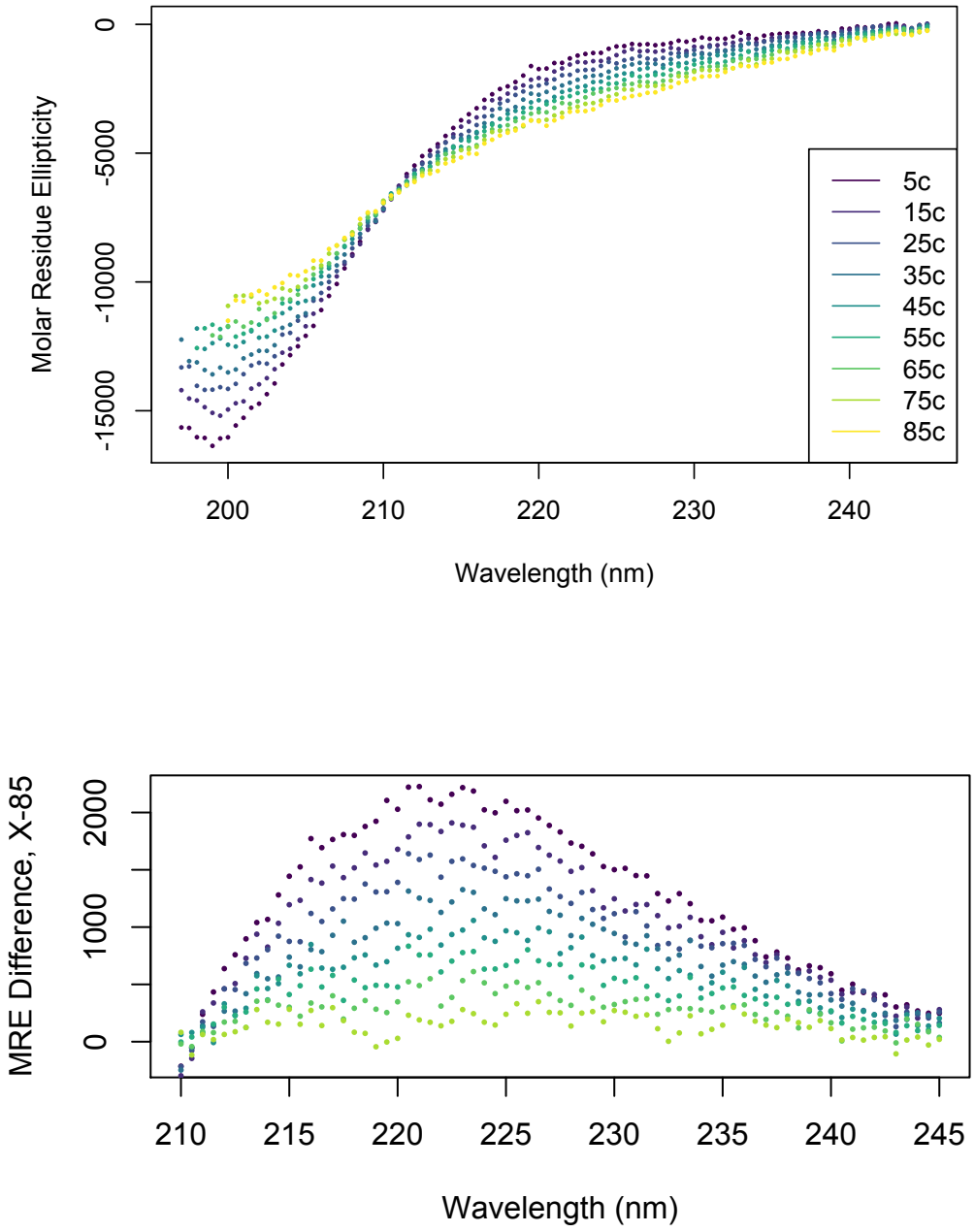


FIG. 26. **Temperature dependent CD spectra for p53(93-1).** Buffer consisted of 10mM sodium phosphate and 100mM sodium chloride at pH 7. *Top;* Molar residual ellipticity (MRE) is measured in units of deg cm^2 $dmol^{-1}$ res^{-1}. Concentration of 17 μM for this peptide. *Bottom;* The difference in MRE for each wavelength between 210 and 240 nm from 85°C spectrum, to show melting of PP_{II} helix.

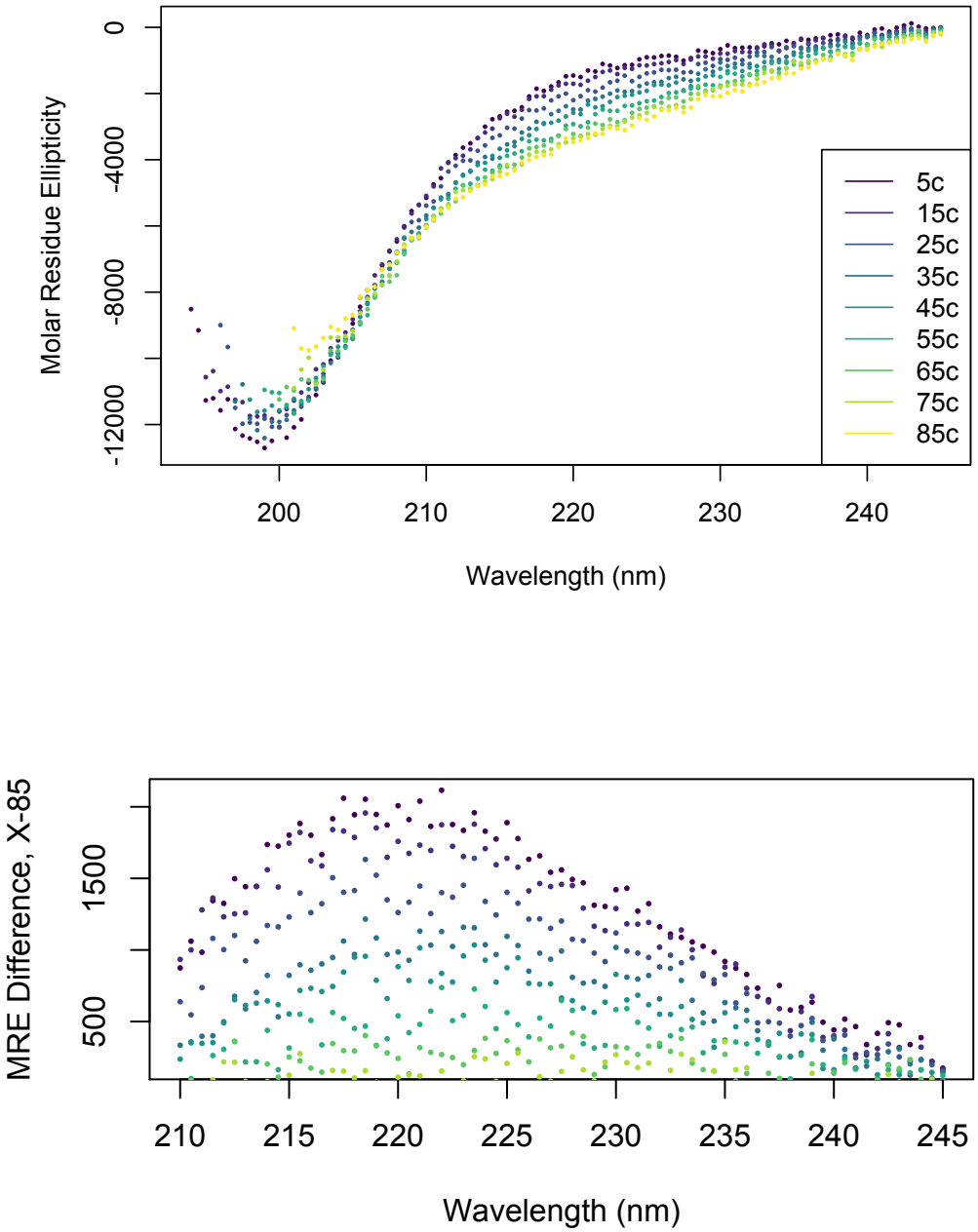
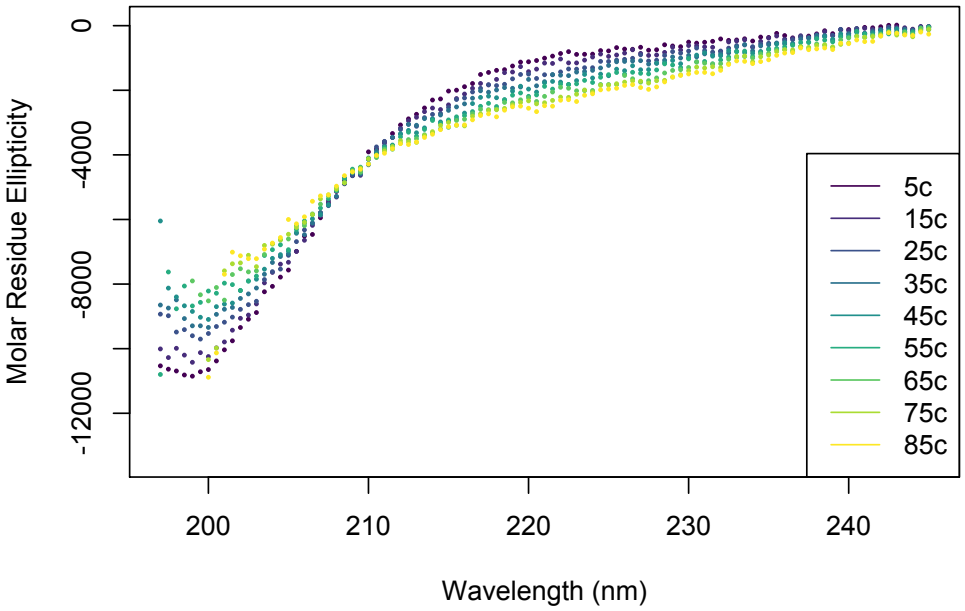


FIG. 27. **Temperature dependent CD spectra for RF11p53.** Buffer consisted of 10mM sodium phosphate and 100mM sodium chloride at pH 7. *Top;* Molar residual ellipticity (MRE) is measured in units of deg cm^2 $dmol^{-1}$ res^{-1}. Concentration of 18 µM for this peptide. *Bottom;* The difference in MRE for each wavelength between 210 and 240 nm from 85°C spectrum, to show melting of PP_{II} helix.



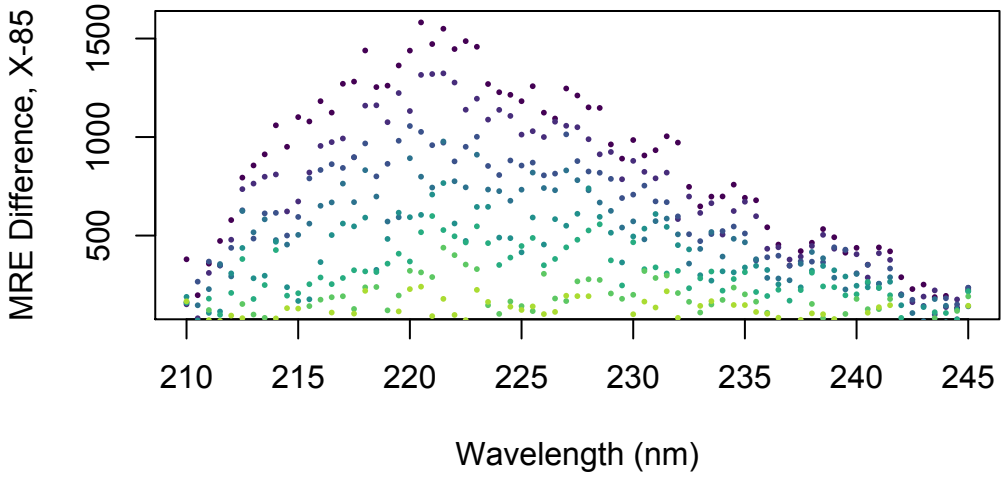


FIG. 28. **Temperature dependent CD spectra for RM11p53.** Buffer consisted of 10mM sodium phosphate and 100mM sodium chloride at pH 7. *Top;* Molar residual ellipticity (MRE) is measured in units of deg cm^2 $dmol^{-1}$ res^{-1}. Concentration of 18 µM for this peptide. *Bottom;* The difference in MRE for each wavelength between 210 and 240 nm from 85°C spectrum, to show melting of PP_{II} helix.

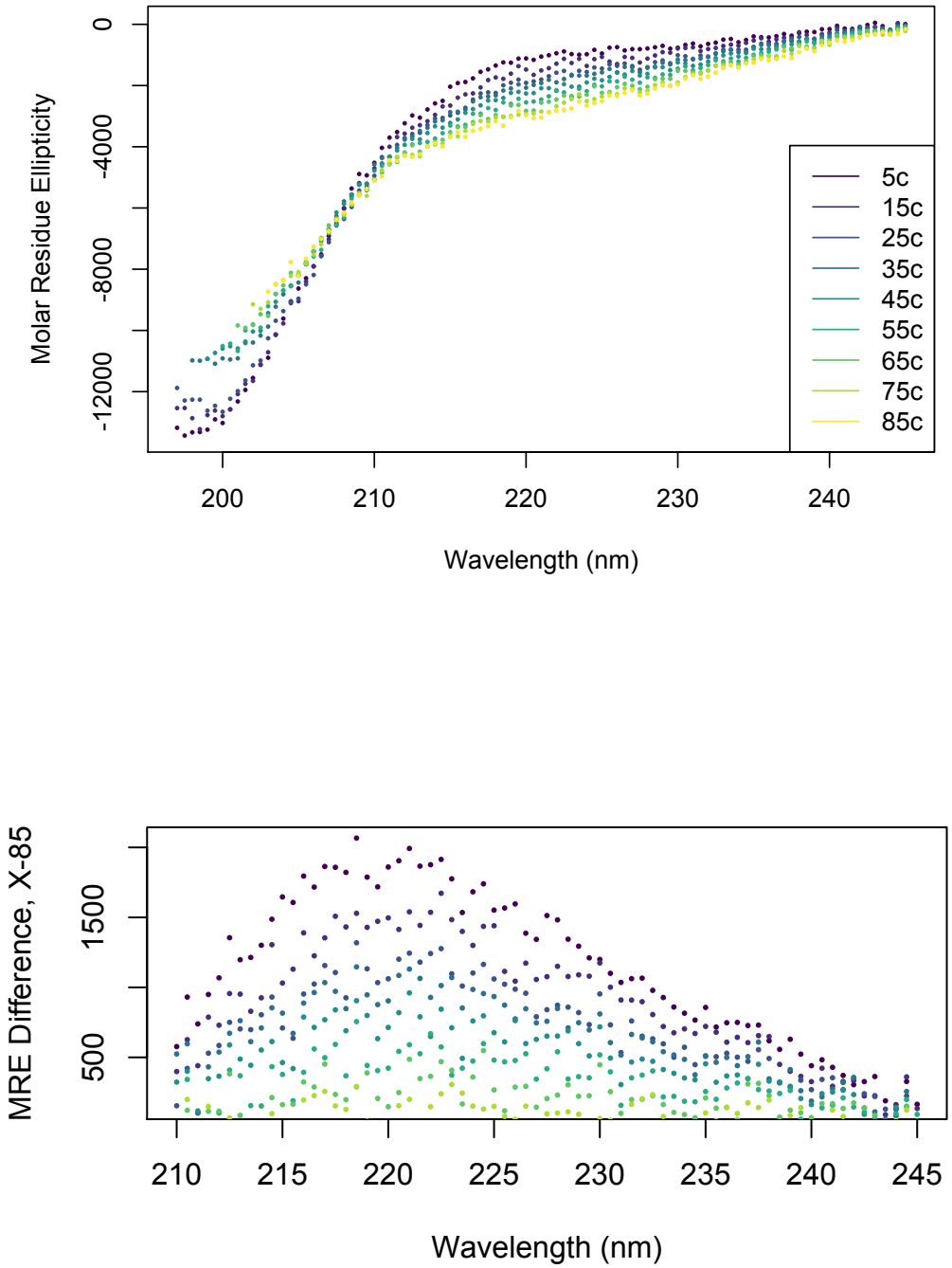


FIG. 29. **Temperature dependent CD spectra for RL11p53.** Buffer consisted of 10mM sodium phosphate and 100mM sodium chloride at pH 7. *Top;* Molar residual ellipticity (MRE) is measured in units of deg cm^2 $dmol^{-1}$ res^{-1}. Concentration of 18 µM for this peptide. *Bottom;* The difference in MRE for each wavelength between 210 and 240 nm from 85°C spectrum, to show melting of PP_{II} helix.

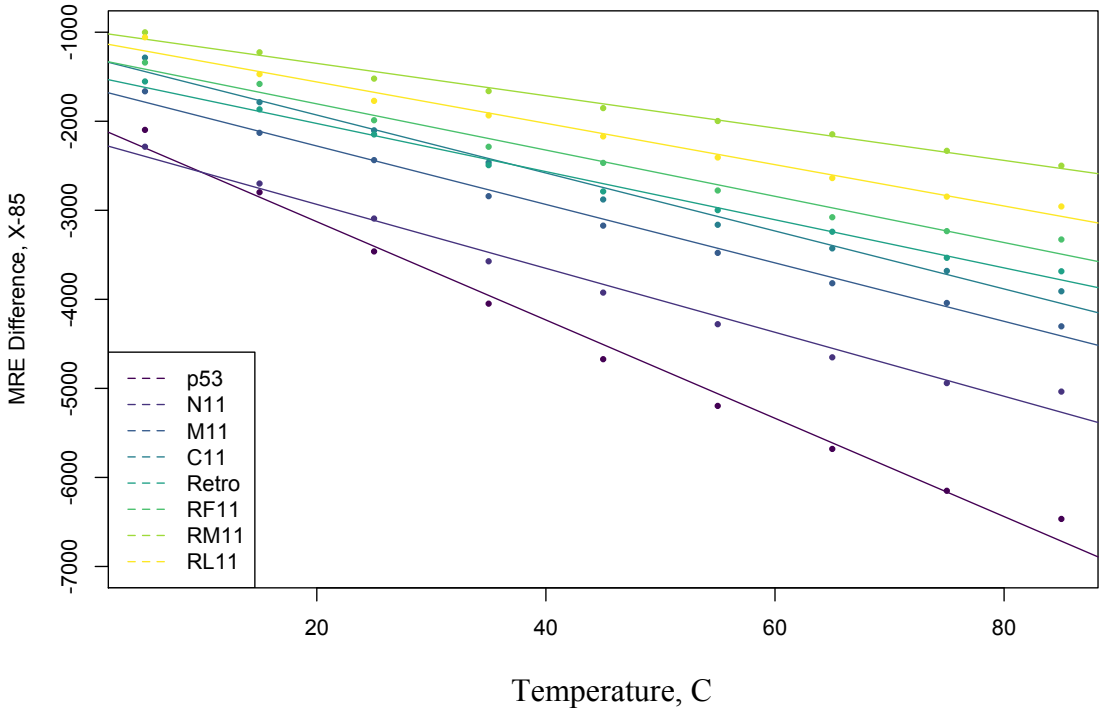


FIG. 30. **Average molar residual ellipticity difference from 85 °C from 220-222 nm.** The linearity of the plots indicates that the collapse of the structure of at this region is noncooperative, as expected for a PP_{II} helix [21].

The structure of the PP_{II} has been shown to represent the most efficient canonical secondary structure for bridging electron transport reactions, which may indicate why disrupting PP_{II} structure at the charge dense region had a greater difference in terms of PP_{II} as measured using CD [46]. The linear relationship of the molar residual ellipticity differences from 85 degrees at 221 nm supports that the hypothesis that the decrease in signal at 221 nm is indicative of melting of the PP_{II} helix. The retro direction variants display lesser PP_{II} as measured by the height of the molar residual ellipticity at 221 nm, seemingly regardless of proline content; the retro wild-type still had less ellipticity at 221 nm than each of the native direction proteins. However, no noticeable change in PP_{II} occurred between variants of the same directionality. These results indicate that changing the direction of the variants reduced PP_{II} structure in the protein variants, but not occurring in substitutions from the same direction.

Size Exclusion Chromatography

An example of a size exclusion chromatogram analysis of p53(93-1) is displayed in **Figure 31**.

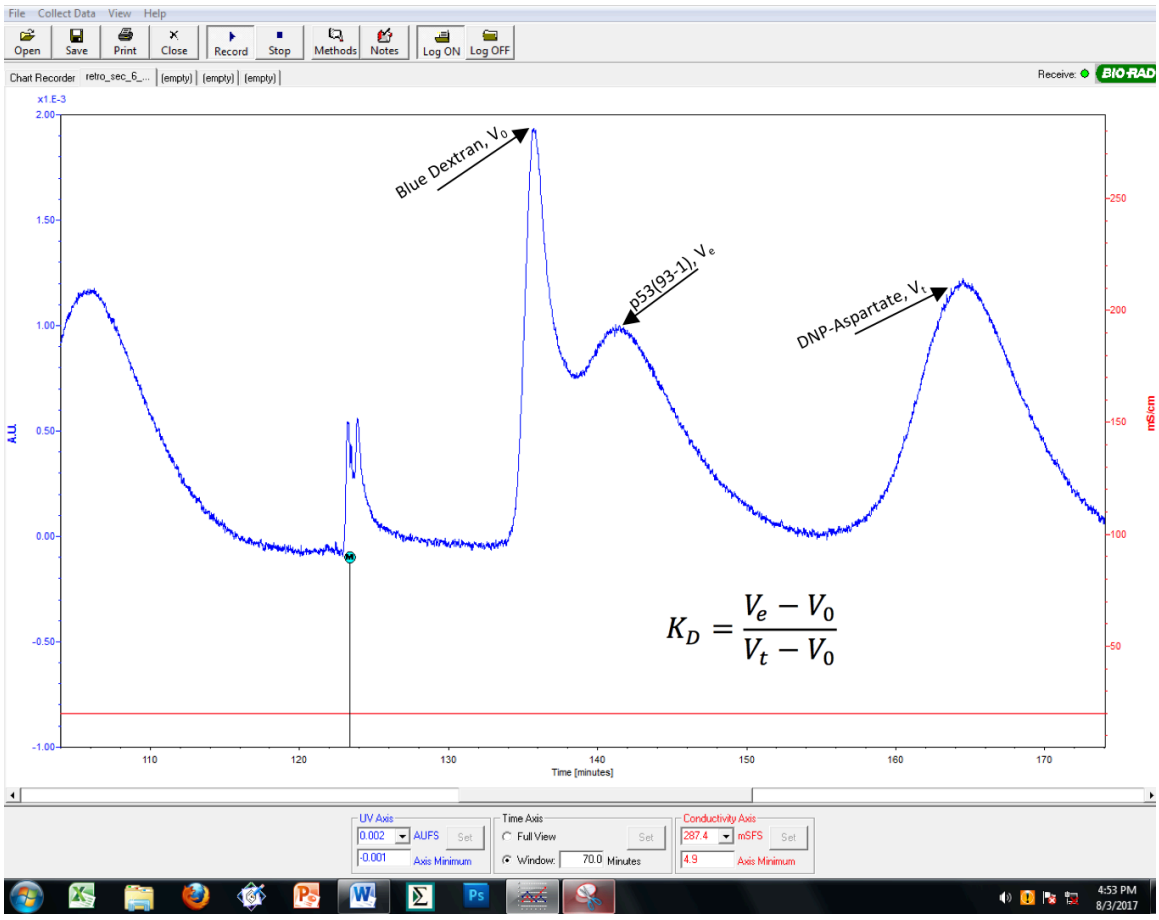


FIG. 31. **Size exclusion chromatogram for p53(93-1).** The peaks indicate measurements of void volume, *Blue Dextran*; elution volume, *p53(93-1)*; total volume, *DNP-Aspartate*.

SEC was performed on three different columns. The first, using G-100 media, measured the R_h for N11p53, M11p53, C11p53, RM11p53, and RL11p53, according to Table 2.

Table 2. **SEC using G-100 media.** R_h calculated from standard curve of K_d and crystallographic R_h, **Figure 32**.

Sample	**N**	**K_d, Average**	**K_d, σ**	**R_h, Å**
N11p53	4	0.2861	0.001	29.68
M11p53	4	0.2887	0.001	29.55
C11p53	4	0.2943	0.003	29.25
RM11p53	4	0.3034	0.005	28.77
RL11p53	5	0.2929	0.006	29.32

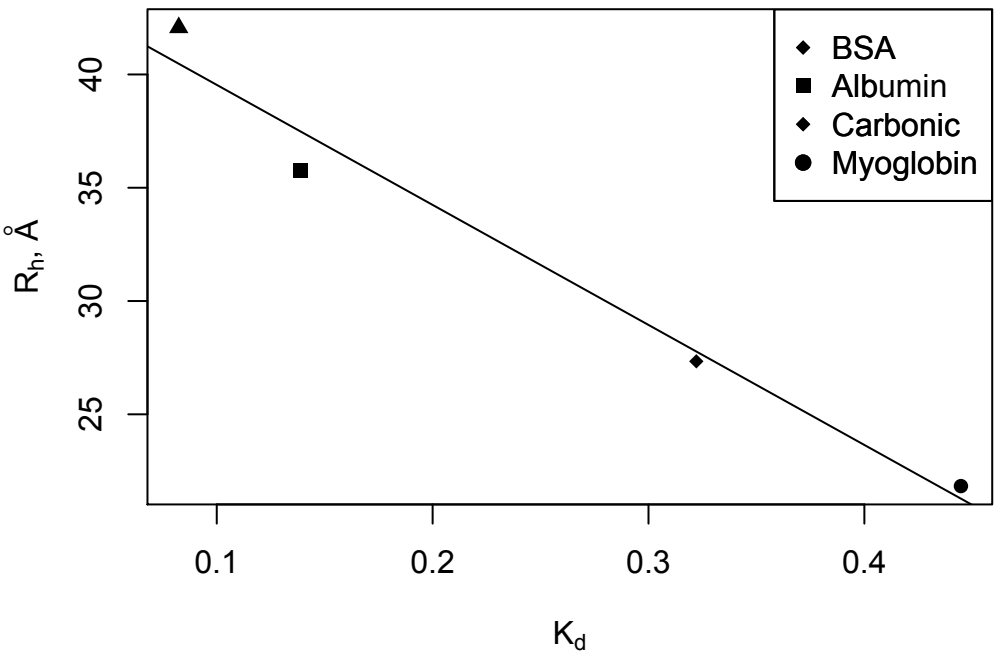


FIG. 32. **Standard curve of SEC proteins on G-100 media.** Linear regression indicates a relationship between K_d and R_h following $R_h = -1.525K_d + 4.9246$, with an $R^2 = 0.959$.

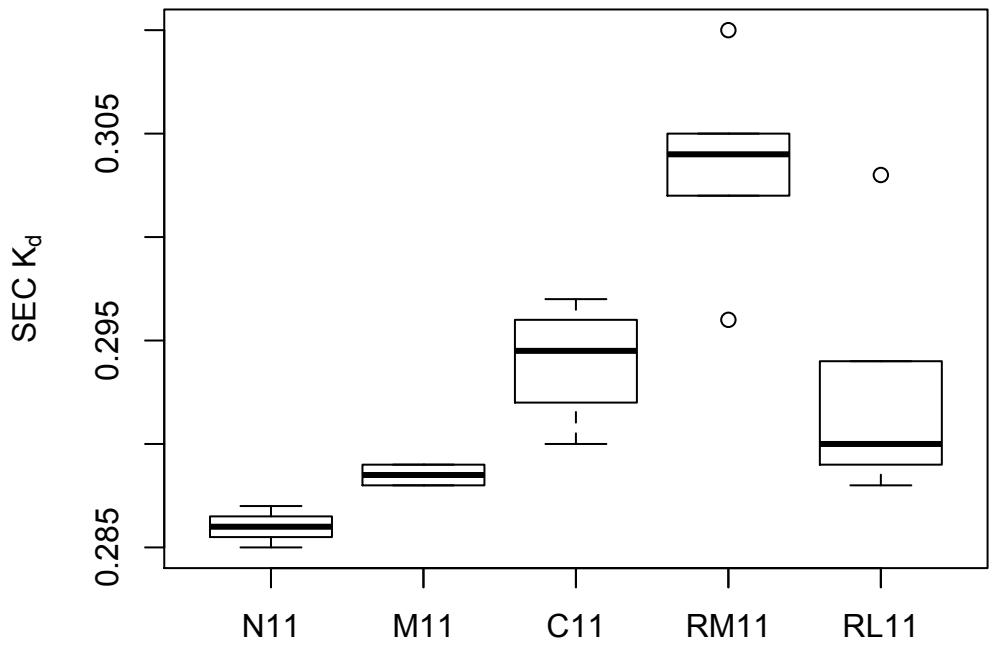


FIG. 33. **Box and whisker plot of SEC results from G-100 media.** Sample sizes; N11, 4; M11, 4; C11, 4; RF11, 4; RL11, 5. The lower and upper limits of the box represent the lower and upper quartile cutoffs, meaning that the regions bounded by the whisker below and above the box represent the lowest and highest quartile of the data, respectively. Outliers are represented as points either below or above 3/2 times the lower or upper quartile, respectively.

Two different SEC runs were obtained on G-75 media. The first compared all proteins, and results are shown in Table 3.

Table 3. **First trial of SEC results on G-75 media.** R_h calculated from standard curve of K_d and crystallographic R_h, **Figure 34**. Two more trials were performed on the p53(1-93) using a purification by another student of the lab, but did not match previous laboratory measurements (laboratory average: 31.81 Å from 3 students; Lance English, Romel Perez, and Erin Tilton).

Sample	**N**	**K_d, Average**	**K_d, σ**	**R_h, Å**
p53(1-93)	1	0.137	-	31.82
N11p53	3	0.205	0.003	29.18
M11p53	2	0.220	0.007	28.61
C11p53	2	0.213	0.011	28.91
p53(93-1)	4	0.177	0.027	30.23
RF11p53	4	0.217	0.004	28.76
RM11p53	4	0.221	0.005	28.59
RL11p53	2	0.224	0.012	28.75

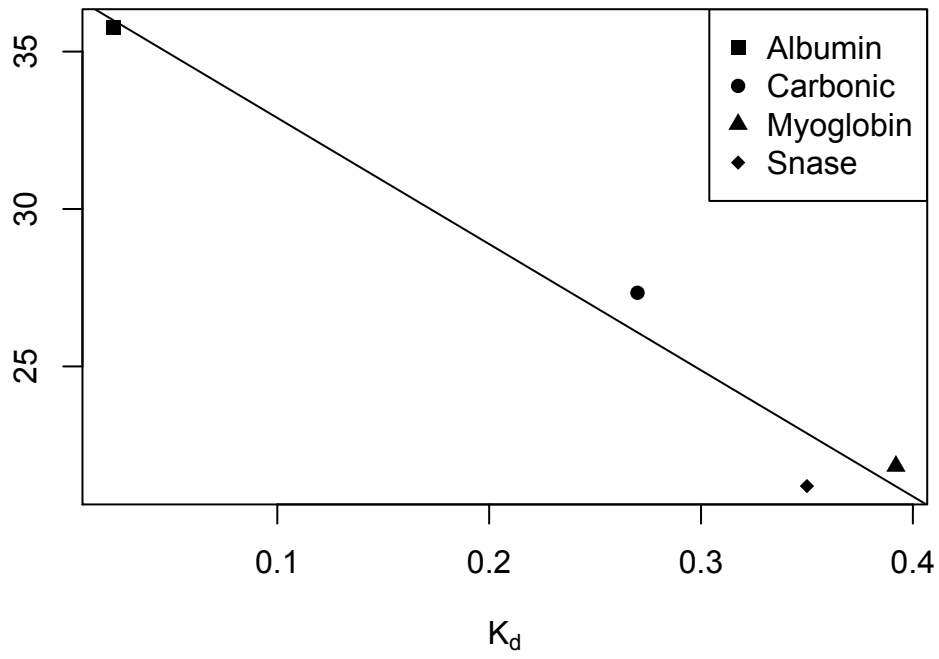


FIG. 34. **Standard curve of SEC proteins on the first trial of G-75 media.** Linear regression indicates a relationship between K_d and R_h following $R_h = -40.046K_d + 36.903$, with an $R^2 = 0.964$. Linear regression without *Staphylococcal* Nuclease: $R_h = -37.163K_d + 36.804$, with an $R^2 = 0.995$.

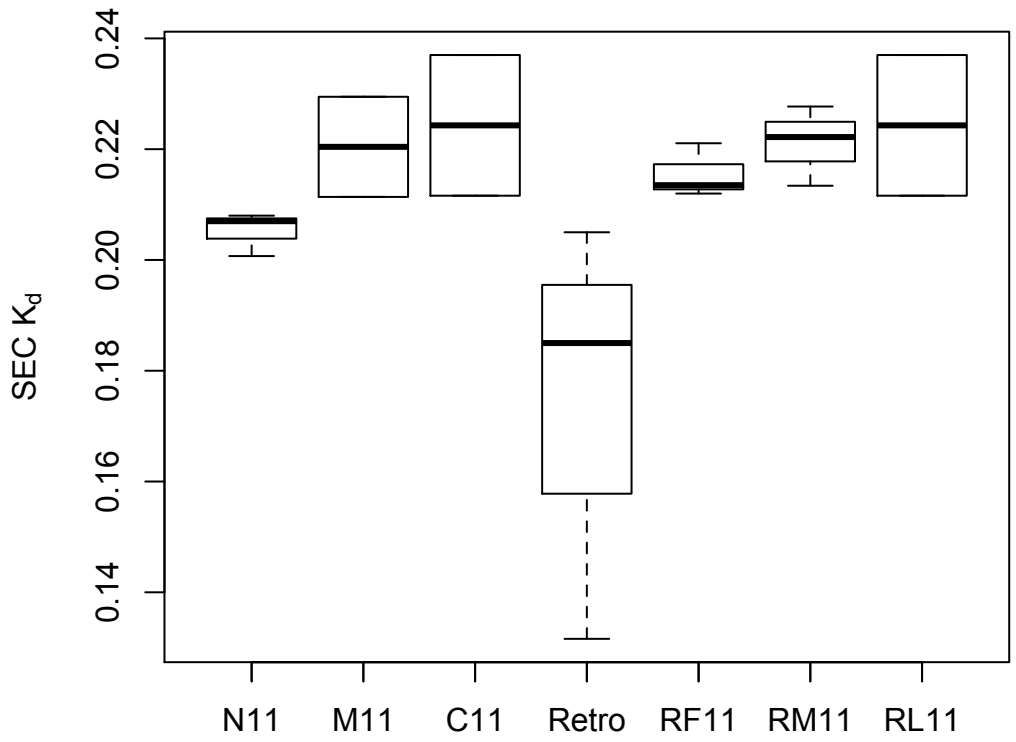


FIG. 35. **Box and whisker plot of SEC results from first trial of G-75 media.** Sample sizes; N11, 3; M11, 2; C11, 2; Retro, 4; RF11, 4; RM11, 3; RL11, 2. The lower and upper limits of the box represent the lower and upper quartile cutoffs, meaning that the regions bounded by the whisker below and above the box represent the lowest and highest quartile of the data, respectively. Outliers are represented as points either below or above 3/2 times the lower or upper quartile, respectively.

The apparent difference of retro p53 and wild type contradicted preliminary experiments in the lab. In order to directly confirm that the wild type and the retro variant were different sizes, as well as measure the RF11p53 variant on another column, a second G-75 column was run, as shown in Table 4.

Table 4. **Second trial of SEC results on G-75 media.** R_h calculated from standard curve of K_d and crystallographic R_h without *Staphylococcal* Nuclease, **Figure 36**.

Sample	**N**	**K_d, Average**	**K_d, σ**	**R_h, Å**
p53(1-93)	6	0.196	0.02	31.93
p53(93-1)	5	0.172	0.02	31.01
RF11p53	4	0.222	0.02	29.43

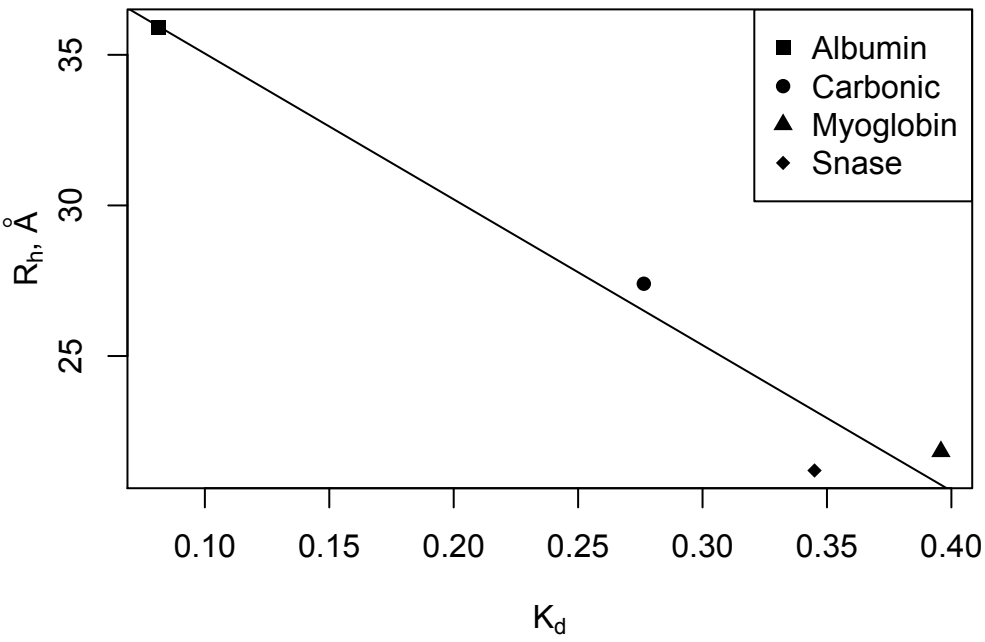


FIG. 36. **Standard curve of SEC proteins on the second trial of G-75 media.** Linear regression indicates a relationship between K_d and R_h following $R_h = -48.435K_d + 36.878$, with an $R^2 = 0.956$. Linear regression without *Staphylococcal* Nuclease: $R_h = -42.15K_d + 38.719$, with an $R^2 = 0.998$.

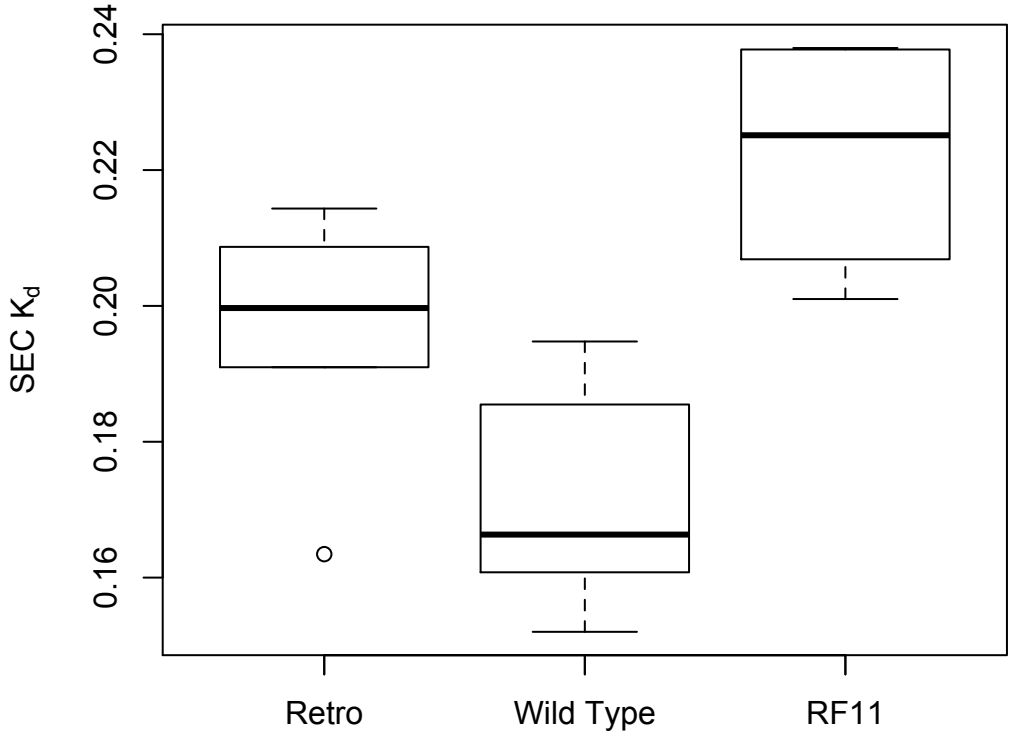


FIG. 37. **Box and whisker plot of SEC results from second trial of G-75 media.** Sample sizes; *Retro*, 6; *Wild Type*, 5, RF11, 4. The lower and upper limits of the box represent the lower and upper quartile cutoffs, meaning that the regions bounded by the whisker below and above the box represent the lowest and highest quartile of the data, respectively. Outliers are represented as points either below or above 3/2 times the lower or upper quartile, respectively.

Size exclusion chromatography (SEC) was performed on each of the purified variants to measure hydrodynamic radius to quantitate differences in intrinsic conformational propensities from various permutations of sequence, depicted in Table 5. For all variants, standardized R_h was decreased in retro sequences relative to their non-retro isomers, indicating the importance of directionality of sequence on the hydrodynamic preferences of p53(1-93). Mutations in the charge rich domain of p53 (N11p53 and RF11p53) were larger than mutations in the other regions, regardless of primary sequence directionality, similar to the trend observed through CD signal peak of Δ220 nm as well as the apparent size through gel electrophoresis.

There was no significant difference between the hydrodynamic radii C11p53 and RL11p53, indicating that changing the proline content (and inductively, PP_{II} propensity) only had significant effects on the hydrodynamic size when altered in an area of high charge (the N11p53/RF11p53 have 13 negatively charged residues between the first and last proline substitutions, and the M11p53/RM11p53 contain 10 negatively charged residues). The existence of unfavorable interactions of charges being ameliorated through the PP_{II} helix has been proposed by various computational studies and may explain the lack of noticeable compaction within the native and retro direction substitutions at the non-charge dense region (C11 and RL11) [25, 47, 48]. Because the PP_{II} may be a mechanism to distribute charges, the N11 and RF11 may be consistently bigger than the other substitutions due to the need to relieve unfavorable charge based interactions through extension of structure through the PP_{II} helix.

The statistical design of the experiment limited the impact of variability of each individual column and allowed for more confident direct comparison. Each protein was run at least for two trials on at least two different columns to ensure consistency of measurement. The average pair-wise difference between any two protein R_h between the three columns was 0.272 Å, representing an error term of 0.92% from the average of all trials (R_h = 29.67 Å), indicating that the column-to-column variation has been sufficiently corrected for in through the measurements of the protein standards. Previously, we established that IDPs occupy a relatively large number of states (average DOS for occupied states is very small, the number of possible states $N >> 1$). In this system, the Central Limit Theorem applies, and that the sampling of distribution of IDPs also follow a normal distribution centered around a mean R_h. The R_h for each peptide was calculated by using the linear regression fit for K_d to calculated R_h standards without *Staphylococcal* Nuclease. *Staphylococcal* Nuclease was not used in determination of the R_h because it 1) consistently underfit the standard curve and 2) was below the detection cutoff for sufficient resolution in the G-100 media. The resulting R_h are shown in Table 5. For comparison, the p53(1-93) protein had been run before by three different students (Lance English, Romel Perez, Erin Tilton) and had an average R_h of 31.82 Å when measured through SEC, compared to 31.91 Å in this experiment.

Table 5. **Calculated R_h from SEC of p53(1-93) variants.** For each column, K_D values were converted to R_h values by protein standard relationships without using *Staphylococcal* Nuclease.

	N	AVERAGE R_H, Å	ST. DEV.
P53(1-93)	6	31.91	0.7
N11P53	7	29.47	0.3
M11P53	6	29.23	0.5
C11P53	6	29.13	0.4
P53(93-1)	10	30.64	0.9
RF11	8	29.09	0.6
RM11	9	28.69	0.2
RL11	7	29.08	0.5

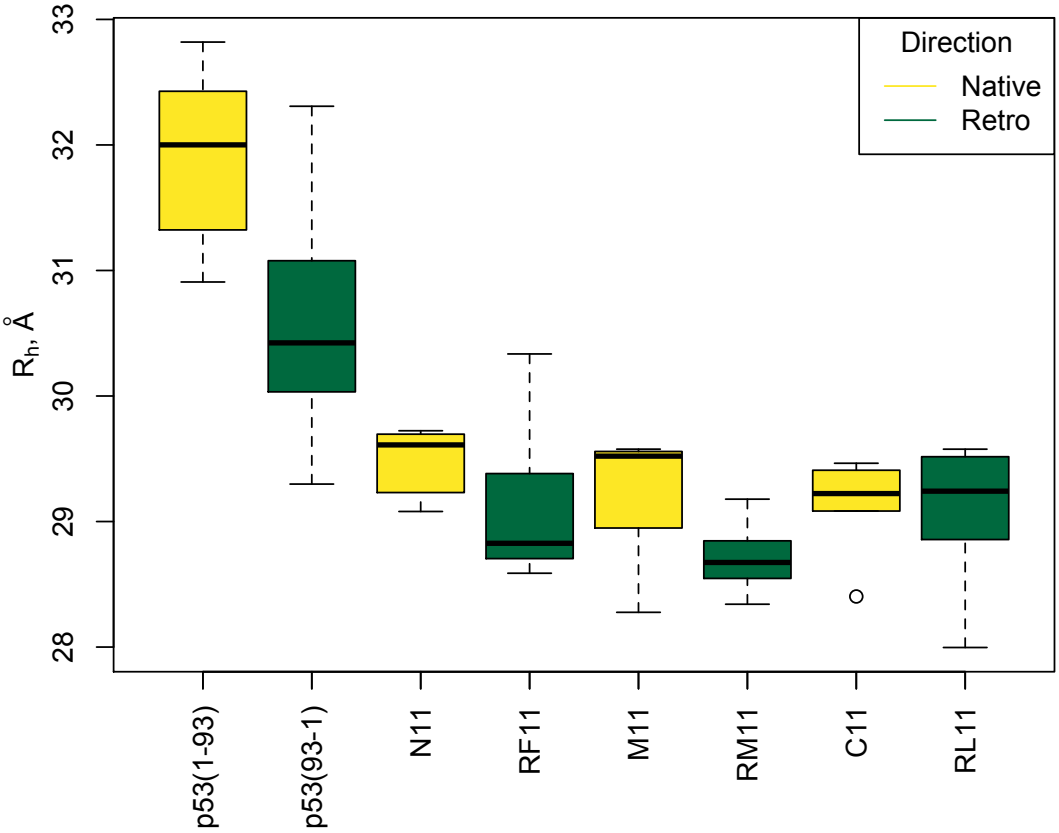


FIG. 38. **Box and whisker plot of calculated R_h of all variants.** R_h is calculated using column standards from protein standard regression without *Staphylococcal* Nuclease. Data from Table 5. The bar located at the center of each box represents the median value. The lower and upper limits of the box represent the lower and upper quartile cutoffs, meaning that the regions bounded by the whisker below and above the box represent the lowest and highest quartile of the data, respectively. Outliers are represented as points either below or above 3/2 times the lower or upper quartile, respectively.

To understand the effects of substitution, a 3x2 ANOVA analysis was performed on the SEC data of the six combinatorically identical variants, as indicated in Table 6-7, using the substation site (columns) and directionality (rows) as factors.

Table 6. **ANOVA design for p53 variants.** *Rows,* direction factor; *Columns*, position of proline to glycine mutations.

	SUBSTITUTION SITE 1	**SUBSTITUTION SITE 2**	**SUBSTITUTION SITE 1**
NORMAL DIRECTION	N11p53	M11p53	C11p53
REVERSE DIRECTION	RF11p53	RM11p53	RL11p53

Table 7. **ANOVA summary table.** *SS*, sum of squares; *df*, degrees of freedom; *MS*, mean of squares; *F*, F-statistic; *P*, p-value.

SOURCE	**SS**	**df**	**MS**	**F**	**P**
ROWS	1.3	1	1.3	6.08	0.0184
COLUMNS	0.98	2	0.49	2.29	0.1154
R X C	0.29	2	0.14	0.68	0.5128
ERROR	7.91	37	0.21		
TOTAL	10.48	42			

The columns represent the different sites of substitution (either the N-terminus, middle of peptide, or C-terminus) and the rows either represent the normal directionality or the reverse directionality. The results indicate that there is a significant difference that can be attributed to the directionality factor ($p = 0.02$) but not for sequence position ($p = 0.12$). There is still a significant ($p < 0.05$) difference between directionality when reducing the compared samples to the first G-75 trial and the G-100 trial, where each protein was run at least for two trials on each, except for RF11, which was run on the first G-75 trial only, indicating that differences are still significant when decreasing the variance in the measurement for RF11p53, which contained the highest standard deviation of the proteins tested.

To further test the hypothesis that directionality a significant effect on the structure of IDPs, were compared using a Welsh's t-test of unequal variances and sample sizes performed on the K_d of p53(1-93) and p53(93-1) from the second G-75 media, giving a p-value of $p = 0.05$. This result indicates a significant pair-wise difference between p53(1-93) and p53(93-1), when run on the same column.

Collectively, both the ANOVA test performed between the six combinatorically equivalent proteins (N11p53, M11p53, C11p53, RF11p53, RM11p53, RL11p53) and the pair-wise Welch's t-test between the retro variant and the wild type indicate a significant directional dependence of R_h, but is not significantly dependent on site of substitution. However, it is important to note that it appears that the substitutions in the region with lower charge density (C11p53 and RL11p53) appear to not differ in directionality. The differences between the native and retro direction variants are displayed in Table 8.

Table 8. **Differences in measured R_h of directional variants through SEC.** The average R_h of p53(1-93) minus the average of p53(93-1), N11p53 minus the average RF11p53, M11p53 minus the average RM11p53, and C11p53 minus the average RL11p53.

	DIFFERENCE, R_H, Å
P53(1-93)	1.27
N11P53	0.37
M11P53	0.55
C11P53	0.05

There is a higher difference in p53(1-93) and its directional variant than any other pair. Almost no difference (0.06 Å) can be observed in the C11p53 and its isomer. These differences indicate that, according to the Flory model of polymer physics, that the reverse in directionality causes less preferential interaction with the solvent, and facilitates an increase in chain-chain interactions. Alternatively, the fractal dimension is reduced due to directionality, indicating the preference of the native directions to be more globally determined, rather than locally determined.

Conclusion

One of the most fundamental ideas in biochemistry is the dependence of a protein's structure on the sequence of amino acids. Anfinsen's "Thermodynamic Hypothesis" states that three-dimensional structure in a given environment is ultimately deterministic to the native state of the lowest Gibbs Free Energy [49]. This was postulated after experiments on ribonuclease A, which returned to its biologically active, native structure after being

completely denatured with urea and having disulfide bridges reduced by 2-mercaptoethanol, but only when allowed to re-oxidize in non-denaturing conditions.

Intrinsically disordered proteins provide a useful system to understand the generalized primary structure of proteins, as they do not occupy a small set of states, but distribute over a larger number of accessible energy states. For folded proteins, primary structure, combined with some sufficiently static environment, eventually leads to a smaller set of structures near some energetic minimum. For IDPs, however, the shift in their *distributions* of states towards a lower energy provide a more general approach for understanding the properties of proteins as it relates to their primary structure.

Reversing the sequence of p53(1-93) dramatically reduced PP_{II} occupancy and R_h regardless of net amino acid content (and any combinatorically based propensity calculation), as indicated by the decrease of signal at 221 nm in the native isomers over the retro isomers. According to the "thermodynamic hypothesis", the protein seeks a lower energy state, and this either implies that 1) the energy cost required to form a PP_{II} helix is increased past the point of being energetically favorable when translated in the opposite direction; 2) another viable conformation becomes more energetically favorable than forming a PP_{II} helix; or potentially both. This finding suggests that determination of PP_{II} helix propensity may be amicable to study through a permutation sensitive model.

Evidence presented here suggests that the hydrodynamic size and PP_{II} propensity depended on the sequence directionality appear to have a charge dependent effect. Because proline causes steric constrictions towards the previous amino acid residue through binding of the side chain to the amine backbone, data was gathered from 1535 unique IDPs in the DISPROT database to compare the number of amino acids that were located proximally to

proline[50, 51]. The number of pre- and post-proline amino acids were compared to the native direction and reverse direction variants of p53(1-93) used in this experiment to obtain an explanation for the collapse of structure indicated by CD and SEC. The percentage of amino acids that occur before and after proline residues in the native and reverse variants of N-terminus p53 and the proteins in the DISPROT database are shown in Table 9. There is an increased prevalence of alanine proximal to proline in p53(1-93), with 36% of prolines being preceded by alanine and 23% of prolines being succeeded by alanine, compared to 11% and 10% of prolines, respectively, in the DISPROT database. The biggest difference in the p53(93-1) relative to p53(1-93) and the prolines in the DISPROT database is the number of prolines after leucine and serine, both of which occur before 14% of all prolines. In the DISPROT database, 6% of prolines are preceded by Leucine, and 9% of prolines are preceded by Serine, while 9% of prolines are preceded by both serine and leucine in p53(1-93). Collectively, this analysis may indicate that the collapse of structure in the retro variants may be attributable to differences in amino acids that precede proline.

Table 9. **Percentage of pre- and post-proline amino acids for p53(1-93) and DISPROT.** 1535 unique intrinsically disordered proteins were analyzed in the DISPROT database. The number of amino acids occurring before and after proline in p53(1-93) is the number that occurs after and before, respectively, in p53(93-1). *None* indicates that proline was the first or last residue in the protein for pre- and post calculations respectively.

	p53(1-93) Pre	*p53(93-1) Pre*	*DISPROT Pre*	*DISPROT Post*
A	36%	23%	11%	10%
C	0%	0%	1%	1%
D	9%	9%	4%	4%
E	9%	9%	8%	9%
F	0%	0%	2%	2%
G	5%	5%	5%	7%
H	0%	0%	2%	1%
I	0%	0%	3%	3%
K	0%	0%	8%	7%
L	9%	14%	6%	5%
M	5%	0%	2%	1%
N	0%	0%	3%	3%
P	9%	9%	10%	10%
Q	0%	5%	6%	6%
R	0%	5%	4%	4%
S	9%	14%	9%	9%
T	5%	5%	6%	8%
V	0%	5%	8%	7%
W	5%	0%	0%	1%
Y	0%	0%	1%	1%
None	0%	0%	1%	1%

Hydrodynamic sizes were significantly altered in retro variants over native direction variants. Altering the positional locations of proline to glycine substitutions did not produce a significant change in hydrodynamic sizes. These results indicate that the dynamic structure of p53, and perhaps of the generalized IDP, is highly dependent on sequence directionality, the extent of which depends highly on the location of charge dense regions.

REFERENCES

1. Onuchic, J.N., Z. Luthey-Schulten, and P.G. Wolynes, *Theory of protein folding: the energy landscape perspective.* Annu Rev Phys Chem, 1997. **48**: p. 545-600.
2. Eisenhaber, F., B. Persson, and P. Argos, *Protein structure prediction: recognition of primary, secondary, and tertiary structural features from amino acid sequence.* Crit Rev Biochem Mol Biol, 1995. **30**(1): p. 1-94.
3. Durbin, S.D. and G. Feher, *Protein crystallization.* Annu Rev Phys Chem, 1996. **47**: p. 171-204.
4. Ward, J.J., et al., *Prediction and functional analysis of native disorder in proteins from the three kingdoms of life.* J Mol Biol, 2004. **337**(3): p. 635-45.
5. Iakoucheva, L.M., et al., *Intrinsic disorder in cell-signaling and cancer-associated proteins.* J Mol Biol, 2002. **323**(3): p. 573-84.
6. Weathers, E.A., et al., *Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein.* FEBS Letters, 2004. **576**: p. 348-352.
7. Uversky, V.N., J.R. Gillespie, and A.L. Fink, *Why are "natively unfolded" proteins unstructured under physiologic conditions?* Proteins, 2000. **41**(3): p. 415-27.
8. Dyson, H.J. and P.E. Wright, *Intrinsically unstructured proteins and their functions.* Nat Rev Mol Cell Biol, 2005. **6**(3): p. 197-208.
9. Li, J., et al., *Thermodynamic dissection of the intrinsically disordered N-terminal domain of human glucocorticoid receptor.* J Biol Chem, 2012. **287**(32): p. 26777-87.
10. Forman-Kay, J.D. and T. Mittag, *From sequence and forces to structure, function, and evolution of intrinsically disordered proteins.* Structure, 2013. **21**(9): p. 1492-9.
11. Mihailescu, M.R. and I.M. Russu, *A signature of the T ---> R transition in human hemoglobin.* Proc Natl Acad Sci U S A, 2001. **98**(7): p. 3773-7.
12. Flory, P.J., *Spatial configuration of macromolecular chains.* Science, 1975. **188**(4195): p. 1268-76.
13. Yu, C., et al., *Structure-based Inhibitor Design for the Intrinsically Disordered Protein c-Myc.* Sci Rep, 2016. **6**: p. 22298.
14. Sormanni, P., F.A. Aprile, and M. Vendruscolo, *Rational design of antibodies targeting specific epitopes within intrinsically disordered proteins.* Proc Natl Acad Sci U S A, 2015. **112**(32): p. 9902-7.
15. Joshi, P. and M. Vendruscolo, *Druggability of Intrinsically Disordered Proteins.* Adv Exp Med Biol, 2015. **870**: p. 383-400.
16. Jeong, C.S. and D. Kim, *Coevolved residues and the functional association for intrinsically disordered proteins.* Pac Symp Biocomput, 2012: p. 140-51.
17. Huang, H. and A. Sarai, *Analysis of the relationships between evolvability, thermodynamics, and the functions of intrinsically disordered proteins/regions.* Comput Biol Chem, 2012. **41**: p. 51-7.
18. Gitlin, L., et al., *Rapid evolution of virus sequences in intrinsically disordered protein regions.* PLoS Pathog, 2014. **10**(12): p. e1004529.
19. Dyson, H.J. and P.E. Wright, *Elucidation of the protein folding landscape by NMR.* Methods Enzymol, 2005. **394**: p. 299-321.

20. Wei, J., et al., *DNA topology confers sequence specificity to nonspecific architectural proteins.* Proc Natl Acad Sci U S A, 2014. **111**(47): p. 16742-7.
21. Chen, K., Z. Liu, and N.R. Kallenbach, *The polyproline II conformation in short alanine peptides is noncooperative.* Proc Natl Acad Sci U S A, 2004. **101**(43): p. 15352-7.
22. Adzhubei, A.A., M.J. Sternberg, and A.A. Makarov, *Polyproline-II helix in proteins: structure and function.* J Mol Biol, 2013. **425**(12): p. 2100-32.
23. Shoulders, M.D. and R.T. Raines, *Collagen structure and stability.* Annu Rev Biochem, 2009. **78**: p. 929-58.
24. Makarov, A.A., et al., *A Conformational study of beta-melanocyte-stimulation hormone.* Biochem Biophys Res Commun, 1975. **67**(4): p. 1378-83.
25. Tomasso, M.E., et al., *Hydrodynamic Radii of Intrinsically Disordered Proteins Determined from Experimental Polyproline II Propensities.* PLoS Comput Biol, 2016. **12**(1): p. e1004686.
26. Zhu, F., et al., *Residual structure in disordered peptides and unfolded proteins from multivariate analysis and ab initio simulation of Raman optical activity data.* Proteins, 2008. **70**(3): p. 823-33.
27. Schweitzer-Stenner, R. and T.J. Measey, *The alanine-rich XAO peptide adopts a heterogeneous population, including turn-like and polyproline II conformations.* Proc Natl Acad Sci U S A, 2007. **104**(16): p. 6649-54.
28. Elam, W.A., et al., *Evolutionary conservation of the polyproline II conformation surrounding intrinsically disordered phosphorylation sites.* Protein Sci, 2013. **22**(4): p. 405-17.
29. English, L.R., et al., *Intrinsic alpha helix propensities compact hydrodynamic radii in intrinsically disordered proteins.* Proteins, 2017. **85**(2): p. 296-311.
30. Perez, R.B., et al., *Alanine and proline content modulate global sensitivity to discrete perturbations in disordered proteins.* Proteins, 2014. **82**(12): p. 3373-84.
31. Wilkins, D.K., et al., *Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques.* Biochemistry, 1999. **38**(50): p. 16424-31.
32. Dewey, T.G., *Fractals in Molecular Biophysics*. 1997: Oxford University Press.
33. Johansen, D., J. Trewhella, and D.P. Goldenberg, *Fractal dimension of an intrinsically disordered protein: small-angle X-ray scattering and computational study of the bacteriophage lambda N protein.* Protein Sci, 2011. **20**(12): p. 1955-70.
34. Mandl, F., *Statistical physics*. 2nd ed. The Manchester physics series. 1988, Chichester West Sussex ; New York: Wiley. xv, 385 p.
35. Uliel, S., A. Fliess, and R. Unger, *Naturally occurring circular permutations in proteins.* Protein Eng, 2001. **14**(8): p. 533-42.
36. Cunningham, B.A., et al., *Favin versus concanavalin A: Circularly permuted amino acid sequences.* Proc Natl Acad Sci U S A, 1979. **76**(7): p. 3218-22.
37. Langridge, T.D., M.J. Tarver, and S.T. Whitten, *Temperature effects on the hydrodynamic radius of the intrinsically disordered N-terminal region of the p53 protein.* Proteins, 2014. **82**(4): p. 668-78.

38. Lopes, J.L., et al., *Distinct circular dichroism spectroscopic signatures of polyproline II and unordered secondary structures: applications in secondary structure analyses.* Protein Sci, 2014. **23**(12): p. 1765-72.
39. Ferreon, J.C. and V.J. Hilser, *The effect of the polyproline II (PPII) conformation on the denatured state entropy.* Protein Sci, 2003. **12**(3): p. 447-57.
40. Surget, S., M.P. Khoury, and J.C. Bourdon, *Uncovering the role of p53 splice variants in human malignancy: a clinical perspective.* Onco Targets Ther, 2013. **7**: p. 57-68.
41. Joerger, A.C. and A.R. Fersht, *The tumor suppressor p53: from structures to drug discovery.* Cold Spring Harb Perspect Biol, 2010. **2**(6): p. a000919.
42. Schaub, L.J., J.C. Campbell, and S.T. Whitten, *Thermal unfolding of the N-terminal region of p53 monitored by circular dichroism spectroscopy.* Protein Sci, 2012. **21**(11): p. 1682-8.
43. Jenkins, L.M., et al., *Two distinct motifs within the p53 transactivation domain bind to the Taz2 domain of p300 and are differentially affected by phosphorylation.* Biochemistry, 2009. **48**(6): p. 1244-55.
44. Jacquet, M.A. and C. Reiss, *In vivo control of promoter and terminator efficiencies at a distance.* Mol Microbiol, 1992. **6**(12): p. 1681-91.
45. Schneider, C.A., W.S. Rasband, and K.W. Eliceiri, *NIH Image to ImageJ: 25 years of image analysis.* Nat Methods, 2012. **9**(7): p. 671-5.
46. Shin, Y.G., M.D. Newton, and S.S. Isied, *Distance dependence of electron transfer across peptides with different secondary structures: the role of Peptide energetics and electronic coupling.* J Am Chem Soc, 2003. **125**(13): p. 3722-32.
47. Krimm, S. and J.E. Mark, *Conformations of polypeptides with ionized side chains of equal length.* Proc Natl Acad Sci U S A, 1968. **60**(4): p. 1122-9.
48. Mao, A.H., et al., *Net charge per residue modulates conformational ensembles of intrinsically disordered proteins.* Proc Natl Acad Sci U S A, 2010. **107**(18): p. 8183-8.
49. Anfinsen, C.B., *Principles that govern the folding of protein chains.* Science, 1973. **181**(4096): p. 223-30.
50. Kovacevic, J.J., *Computational analysis of position-dependent disorder content in DisProt database.* Genomics Proteomics Bioinformatics, 2012. **10**(3): p. 158-65.
51. Piovesan, D., et al., *DisProt 7.0: a major update of the database of disordered proteins.* Nucleic Acids Res, 2017. **45**(D1): p. D219-D227.