Intrinsically Disordered Protein Polymer Libraries

as Tools to Understand Protein Hydrophobicity

by

Nicholas Tang

Department of Biomedical Engineering
Duke University

Date:_____
Approved:

_____
Ashutosh Chilkoti, Supervisor

_____
Lingchong You

_____
Charles Gersbach

_____
Michael Lynch

_____
Sayan Mukherjee

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor
of Philosophy in the Department of
Biomedical Engineering in the Graduate School
of Duke University

2019

ABSTRACT

Intrinsically Disordered Protein Polymer Libraries

as Tools to Understand Protein Hydrophobicity

by

Nicholas Tang

Department of Biomedical Engineering
Duke University

Date:_____
Approved:

_____
Ashutosh Chilkoti, Supervisor

_____
Lingchong You

_____
Charles Gersbach

_____
Michael Lynch

_____
Sayan Mukherjee

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biomedical Engineering in the Graduate School of Duke University

2019

Copyright by
Nicholas Tang
2019

Abstract

Intrinsically disordered protein polymers (IDPPs) are repetitive biopolymers that, when enriched with prolines, glycines, and aliphatic amino acids, have observable lower critical solution temperature (LCST) phase transition behavior at physiologically relevant temperature and concentration ranges. This behavior is a striking feature of disordered proteins in nature, where chemical or physical stimuli lead to sharp conformational or phase transitions. Accordingly, protein-based polymers have been designed to mimic these behaviors, leading to a broad range of biotechnological applications. This work is driven by two approaches. In our science focused approach, we developed a polymer-physics based framework for understanding IDPP hydrophobicity using the relationship between phase transition temperature and globule surface tension. This physics-based framework has allowed us to better understand the unified contributions of chain length, concentration, temperature, and individual amino acid side chains to IDPP hydrophobicity by studying phase transition data. In our engineering focused approach, we developed novel tools that enable the high throughput discovery of new proteins that exhibit phase transitions, in order to increase the number of known stimuli responsive peptide sequence motifs beyond the limits of bioinspired design. The exhaustive discovery of new proteins that exhibit phase transitions consists of gene synthesis and protein screening. We developed two key

technologies that has enabled (1) the scalable synthesis of repetitive gene libraries using a novel graph theoretic gene optimization approach (Codon Scrambling) and (2) the pooled synthesis of large complex gene libraries from libraries of oligonucleotides. Combined with pipelines for the screening of phase transition behavior, these technologies have enabled us to generate a diverse library of protein sequences necessary to validate our theoretical models. Finally, we developed an algorithm for the *de novo* design of nonrepetitive protein sequences that exhibit phase transition behavior, further broadening the sequence space of stimuli responsive synthetic IDPPs.

Contents

Abstract iv

List of Tables ix

List of Figures x

Acknowledgements xiii

1. Introduction 1

1.1 Protein hydrophobicity 1

1.2 Intrinsically disordered protein polymers (IDPPs) are ideal model proteins for hydrophobicity 2

1.3 LCST phase transition behavior in ELPs 3

2. Towards an equation for protein hydrophobicity 7

2.1 Empirical models for phase transition 7

2.2 Development of an equation for the binodal curve for dilute solutions 8

2.3 Sequence dependence of the equation for an LCST binodal using experimental observations of a family of ELPs 11

2.4 Observations on θ and θ_0 for a more general set of IDPPs 21

3. Codon scrambling: an enabling technology 31

3.1 Introduction 32

3.2 Formulation of the codon scrambling problem 36

3.3 Development of the codon scrambling problem and its MILP formulation 39

3.4 Computational analysis for optimization of repetitive proteins 46

3.5 Codon scrambling enables gene assembly with PCR 50

3.6 Codon scrambling enables Golden Gate and Gibson Assembly 52

3.7 Effects of codon scrambling on recombinant protein expression 55

3.8 Materials and methods 62

3.8.1 Gene design and assembly with PCA 62

3.8.2 PCR and Golden Gate assembly of ELP 64

3.8.3 ELP protein expression and purification 65

3.8.4 PCR and Gibson assembly 66

3.8.5 Recombinant protein expression 67

3.8.6 Computation 67

4. Gene libraries for the discovery of novel protein polymers 68

4.1 Introduction 68

4.2 Oligonucleotide libraries as materials for gene synthesis 70

4.2 RCA-based pooled polymer library synthesis 71

4.3 Next generation sequencing verifies library diversity 77

4.4 Colony filtration screens for soluble expression 79

4.5 Characterization of soluble proteins 85

4.6 Materials and methods 87

4.6.1 Gene library synthesis using RCA 87

4.6.2 NGS analysis of gene library 89

4.6.3 Colony screening for solubility and expression 89

4.6.4 Protein expression and purification 91

4.6.5 Cloud point characterization 92

5. Additional tools for peptide polymer production 93

5.1 SYPRO orange assay for the high throughput determination of phase behavior . 93

5.2 Unstructured non-repetitive polypeptides have LCST behavior 97

5.3 Materials and methods 105

6. Validation of the phase transition model and future directions 106

7. Conclusions 114

Appendix 120

References 148

Biography 159

List of Tables

Table 1: MILP formulation 38

Table 2: Computational results for the optimization of repetitive proteins 47

Table 3: Gene synthesis error rates 50

Table 4: Rounds of cloning to assemble full-length genes 54

Table 5: Approximate costs for gene library synthesis 72

Table 6: Summary of sanger sequencing results 81

Table 7: MILP formulation of the constraint satisfaction program 100

Table 8: Characterization of the repetitiveness of repetitive and non-repetitive polypeptide sequences 105

Table 9: Purified protein polymer sequences. 112

Table 10: Motifs of 1020 microarray-synthesized oligonucleotides 120

Table 11: Designed nucleotide sequences of 1020 microarray-synthesized oligonucleotides 121

Table 12: Summary of alignments of next generation sequencing (NGS) data grouped by oligonucleotide reference ID and concise idiosyncratic gapped alignment report (CIGAR) 126

Table 13: Summary of sanger sequencing of unique screened colonies 130

Table 14: Nonrepetitive sequences that display LCST behavior 137

Table 15: Nucleotide sequences of genes generated using codon scrambling for high throughput DNA synthesis and protein production in *E. coli* 138

Table 16: Summary of published IDPP phase transition data from various sources 145

List of Figures

Figure 1: Schematic phase diagram for several elastin like polypeptides based on the binodal curve for approximately 120 repeats of GVGVP. 5

Figure 2: Verification of (A) Φ'' and (B) ξ_T/b. 15

Figure 3: $1/c_1$ and $1/\theta$ are coupled and linearly related. 15

Figure 4: Fitted values for 3.14 c_1 versus fraction of alanine vary in a near linear manner with increasing substitution of valines to alanines. 17

Figure 5: Verification of R_g and γ/kT. 19

Figure 6: Temperature dependence of the length-collapsed miscibility gap. 20

Figure 7: Visualization of the collapse of data. 21

Figure 8: Linearization of the equation for T for independent fits for a variety of different sequence patterns. 23

Figure 9: Linear correlation between fitted c_1 and $c_{1,\text{linear}}$ from a multivariate-linear model. 25

Figure 10: Lack of correlation between θ and θ_0. 26

Figure 11: Trajectories in the $\theta/(\theta - \theta_0) - \theta_0$ space for increasing pH (A) and the nonlinear effects of pH on θ_0 (B) for histidine-containing sequences. 28

Figure 12: The trajectory in $\theta/(\theta - \theta_0) - \theta_0$ space for increasing salt (A) and the linear effects of NaCl concentration on θ_0 (B) for V[0.2]A[0.8]. 29

Figure 13: Example diagram of a multipartite n-mer codon De Bruijn subgraph with a window length $k = 1$ and a motif length $r = 5$ (GVGVP motif). 35

Figure 14: Nucleotide sequence dot plot. 48

Figure 15: Timings of CPLEX optimization. 49

Figure 16: Gene assembly of a diverse set of repetitive proteins. 51

Figure 17: Golden Gate assembly of ELP[V-150]. 53

Figure 18: Microfluidic electropherograms of PCR products of optimized genes. 55

Figure 19: Protein expression of a diverse set of repetitive proteins. 56

Figure 20: Effects of codon scrambling on protein expression and purification..... 57

Figure 21: Expression and utilitization of repetitive proteins. 60

Figure 22: Purification of ELPs..... 61

Figure 23: Generation of repetitive IDP libraries..... 73

Figure 24: Restriction digestion of RCA products generated with 0% 5-methyl-dCTP (left) and 10% 5-methyl-dCTP (right)..... 75

Figure 25: Partial digestion of RCA products from 36-nt oligonucleotide material (left) and 72-nt oligonucleotide material (right). 77

Figure 26: Illumina Miseq analysis of cloned library of plasmid DNA verifies the presence of at least 900 unique repetitive amino acid sequences. 79

Figure 27: The CoFi method is used to identify and locate *E. coli* colonies expressing soluble proteins. 80

Figure 28: PCR directly from library-transformed *E. coli* colonies on dense plates visualized on agarose gel electrophoresis. 81

Figure 29: Examples of (a) dot blot and PAGE analysis of (b) soluble and (c) insoluble lysate fractions to determine target protein solubility..... 84

Figure 30: Protein polymers lack ordered secondary structures as shown by the CD spectra at 25 °C of a subset of soluble protein polymers..... 85

Figure 31: Turbidimetry data for protein polymers that exhibit LCST behavior between 15°C and 85°C in PBS, or PBS plus Urea for [GGMLAP]24*. 86

Figure 32: SYPRO orange assay with various lengths of elastin like polypeptides (ELPs). 95

Figure 33: SYPRO orange assay with ELPs. 96

Figure 34: Amino acid compositions of the nonrepetitive and repetitive versions of the A[0.5] and A[0.2] sequences. 101

Figure 35: Characterization of repetitive and nonrepetitive versions of A[0.2] and A[0.5] comprising (a) 200 amino acids and (b) 400 amino acids. 102

Figure 36: Amino acid compositions of exemplary non-repetitive unstructured polypeptides. 103

Figure 37: Transition temperature characterization of exemplary non-repetitive unstructured polypeptides 104

Figure 38: (A) Summary of ~60 newly produced sequences and of ~190 previously published repetitive proline and glycine-rich sequences 109

Figure 39: Linear prediction of fitted θ using Urry hydropathy. 111

Figure 40: Comparison of measured T_t and predicted T_t using the LCST IDPP phase transition model with estimated values of θ and θ_0 = 272.63. 112

Figure 41: Distribution of amino acids encoded by degenerate codons, DYD and RRR. 116

Figure 42: Urry's relationship between transition T_t and change transition heat upon amino acid substitution near θ. 118

Acknowledgements

This work was supported through grants to Dr. Ashutosh Chilkoti from the National Institutes of Health. I thank the Center for Biomedical and Tissue Engineering for supporting me with a graduate student training grant from the National Institutes of Health. The following people and Duke facilities have provided important resources for this study: Nicolas Devos and the Duke GCB NGS Core, Steven Slater and the Duke DHVI Core, and Heather Hemric and the Duke Microarray Core.

I am grateful to all members of my dissertation committee for their time and guidance throughout the course of this work. I would like to especially thank Dr. Ashutosh Chilkoti for his years of guidance on a wide range of topics and expertise. Our discussions have encouraged and shaped my view of science and engineering. Dr. Michael Rubinstein has given me significant scientific advice for polymer physics and trained me to approach polymer science with rigor. I also thank Dr. Sayan Mukherjee for introducing me to the world of mathematical optimization theory.

Garrett Kelly has directly aided me in the creation and optimization of protein libraries and non-repetitive IDPPs. His scientific efforts have been critical towards the last few years of this work. I would like to thank Dr. Liel Sapir, as our joint discussions with Dr. Michael Rubinstein during the last few months of this work have been essential towards shaping the physics model. I also thank Michael Dzuricky for scientific discussions and encouragement which were very influential to the progress of this work.

I would like to thank Drs. Wafa Hassouneh and Felipe Garcia-Quiroz for their guidance during the early years of this project and their hours of discussions. I would like to thank the rest of the Chilkoti lab for their generous years of troubleshooting and intellectual help.

Finally, I would like to thank my father, mother and brother for their patience and encouragement that has led me to this point in my academic career.

1. Introduction

1.1 Protein hydrophobicity

In biology, hydrophobicity is an important driving force for protein folding, stability, binding interactions, as well as a number of macromolecular assemblies such as lipid bilayers and viruses. Water molecules surrounding nonpolar molecules incur a free energy cost, which is described as the "hydrophobic effect". The water molecules arrange themselves in an ordered and cohesive clathrate-like structure, which then causes the nonpolar molecules to shrink and reduce its interfacial area with water in order to reduce entropy. This is because the attractive forces between water molecules is much stronger than the van der Waals forces between polar groups of polymers and water (*1*). A useful physical interpretation of hydrophobicity at liquid interfaces is surface tension, the energy per unit area that causes this shrinking. The cost to hydrate a spherical cavity of radius R is $\Delta G \approx 4\pi R^2\gamma + (4/3)\pi R^3 p$, where γ is the liquid-vapour surface tension, and p is the pressure (*2*).

Protein hydrophobicity in water depends on solvent accessibility and sequence content, as well as other parameters such as size, concentration, and temperature. Due to its complexity, protein hydrophobicity is often correlated with sequence information empirically using hydrophobicity scales. This relies on the observation that each amino acid can be delineated on a range from hydrophobic to hydrophilic, and that the sum of their independent contributions to a sequence or subsequence can be used to

characterize or identify protein domains. There are over 150 hydrophobicity scales mostly derived from experimental measurements of native proteins (*3*), many of which differ substantially from each other- a testament to the difficulty of measuring fundamental properties from complex proteins. Many of the successful attempts of using hydrophobicity scales to predict or classify proteins based on structure and/or function involve feature selection of multiple hydrophobicity scales and heuristic models based on evolutionary approaches or neural networks (*4*, *5*). These methods are focused on predictive approaches.

1.2 Intrinsically disordered protein polymers (IDPPs) are ideal model proteins for hydrophobicity

From a reductionist approach, to further improve estimations of protein hydrophobicity, an equation from physical principles would be valuable, specifically one that explains the contributions of amino acids in the context of factors such as size, concentration, temperature and solvent quality. As a first step, a model protein must be chosen to reduce the complexities of protein structure. It must: (1) be devoid of stable secondary structures (intrinsically disordered) which promote compact water-excluding structures, (2) remain hydrated upon collapse from coil to globule, and (3) have repetitive, low-complexity sequence content in order to modeled as homopolymers. With these conditions, all side chains can interface with water and therefore contribute to surface tension. We identify a class of synthetic IDPPs based on the sequence-features of elastin are the ideal model protein which meets the criteria: (*6*). Its most important

sequence defining characteristic are repeating ~5 amino acid motifs which contain one proline and at least one glycine, ensuring low complexity and periodically-distributed prolines and glycines (*7*).

Elastin proteins, including the 60 kDa tropoelastin molecule, have remarkably unique protein-properties including high resilience and extensibility. This may be attributed to the high proline and glycine content, which ensure they remain hydrated and disordered so that they readily undergo extension and elastic recoil (*7*, *8*). Recent work has described its secondary structure as fluctuating β-turns, with a lifetime of <1 ns, consisting mostly of the tetrapeptide VPGV, based on simulation work (*8*, *9*) and spectroscopic data (*10*, *11*). The sparse and highly transient secondary structure, and persistent hydration is largely explained by a combination of proline's fixed φ dihedral angle and glycine's conformational entropy.

1.3 LCST phase transition behavior in ELPs

Bovine tropoelastin contains 11 regular repeats of the pentapeptide motif VPGVG which was found to exhibit lower critical solution temperature (LCST) behavior, also termed inverse transition temperature (ITT). LCST behavior was first observed in over 70 synthetic polymers such as poly(N-isopropyl acrylamide) (*1*, *12*). In poor solvents, polymers exhibiting LCST behavior exhibit temperature dependent coil to globule collapse at a cloud point temperature, also termed transition temperature. Below this temperature, polymer solutions are miscible or soluble. Above it, they are partially

miscible, separating into a globule and sediment phase. At equilibrium, a small fraction of globules remains in solution with concentration, Φ', because their surface energy is balanced by their translational entropy. Due to the energetic cost due to surface tension, the rest of the globules form large polymer-rich clusters with higher concentration, Φ''. Overall, the composition-temperature relationship can be described by a liquid-liquid phase diagram, where the one-phase and two-phase regions can be separated by a solubility boundary called the binodal curve. The binodal for binary mixtures coincides with the coexistence curve, because the two compositions that coexist can be read off the binodal. This curve is concave-up where the lowest point corresponds to the critical point (T_c,Φ^*). For $\Phi<\Phi^*$, the solution is called dilute, where the average distance between chains is larger than their size. For $\Phi>\Phi^*$, the solution is called semidilute, where the polymer coils overlap and dominate the physical properties of the solution.

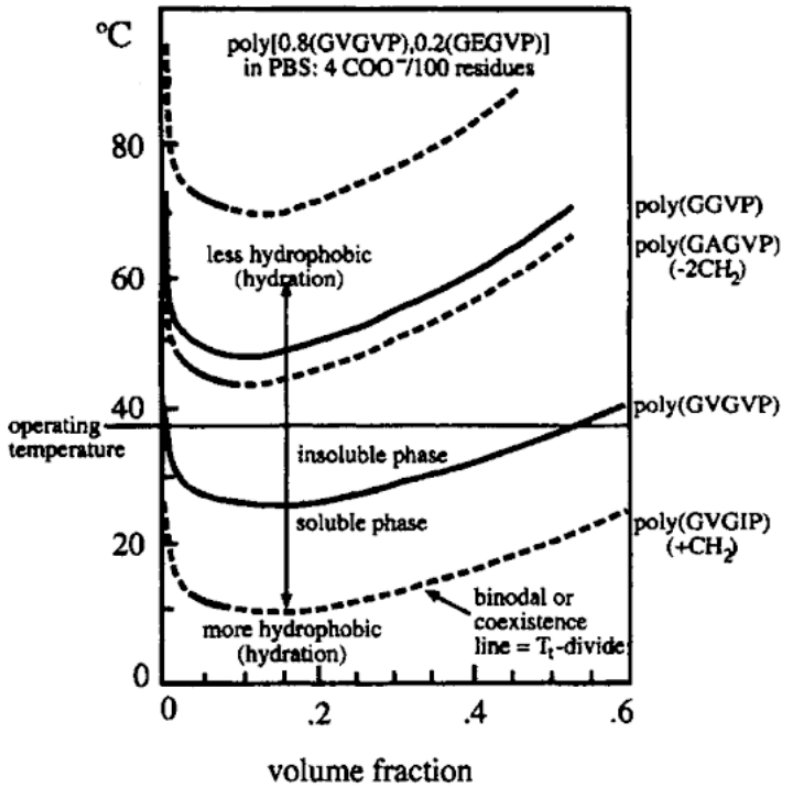

Figure 1: Schematic phase diagram for several elastin like polypeptides based on the binodal curve for approximately 120 repeats of GVGVP. Adapted from (*13*)**.**

Bioinspired by the VPGVG motif of the hydrophobic domain of tropoelastin, Urry and coworkers used chemical synthesis to produce high molecular variants of poly[VPGVG]. They found that partial substitution of sequences of the form poly[$(VPGVG)_v(VPGXG)_x$] also exhibited LCST behavior, where v and x are mole fractions $v+x = 1$, and X is any guest residue except for proline. Cloud points were measured by measuring turbidity formation, at high molecular weights (> 50 kDa) and concentrations (40 mg/mL). Importantly, linearity was observed between T_t and mole

fraction x for many residues, suggesting that LCST can be tuned by sequence content. This class of synthetic proteins is termed elastin like polypeptides (ELPs).

The phase transition of ELPs have been resolved as a mixture of endothermic and exothermic components by temperature modulated differential scanning calorimetry (TMDSC). The exothermic component is due to van der Waals cohesive interactions while the endothermic component is due to loss of hydrophobic hydration. The endothermic component is more than three greater than the exothermic component, as required for phase transition to occur (*14*).

2. Towards an equation for protein hydrophobicity

The following chapter contains work which will be published in a peer-reviewed journal.

2.1 Empirical models for phase transition

Chilkoti's group produced ELPs using recombinant synthesis in *Escherichia coli* rather than chemical synthesis. As the genes encoding ELPs are constructed through recombinant DNA techniques, precise chain length and residue compositions can be attained. This allowed Meyer and Chilkoti to develop an empirical equation quantifying effects of length and concentration on cloud point temperature for dilute solutions for several variants of ELP,

$$T_t = T_{t,c} + \frac{k}{Length}\ln(\frac{C_c}{Conc}) \quad \textbf{(1)}$$

where $T_{t,c}$, k, and C_c are independently estimated parameters, *Length* is chain length, *Conc* is molar concentration, and T_t is transition, or cloud point, temperature (*15*). The equation was further expanded to predict the T_t for a family of ELPs poly[$(VPGVG)_v(VPGAG)_x$] where v and x are mole fractions of VPGVG and VPGAG respectively:

$$T_t = Ae^{Bf} + \frac{Ce^{Df}}{Length}\ln(\frac{Ee^{Ff}}{Conc}) \quad \textbf{(2)}$$

where *A*-*F* are fitted parameters and *f* is the ratio of alanines to the total number of guest residue positions (*16*). Compared to polymers by Urry and coworkers, the distribution

of VPGVG and VPGAG is regular instead of statistical, as the genes encoding the ELPs were constructed from nucleotide oligomer units encoding 10 well-mixed pentamers of VPGVG and VPGAG. The role of pH was also modeled for ELPs containing acidic (Glu) or basic (His) residues by substituting Meyer and Chilkoti's equation for T_{pro} and T_{depro} in the following equation based on the Henderson-Hasselbach relationship:

$$T_t = T_{pro} + \frac{T_{depro} - T_{pro}}{1 + 10^{(pKa - pH)}} \quad \textbf{(3)}$$

where T_{depro} is the transition temperature of a fully deprotonated ELP at a high pH, and T_{pro} is the transition temperature of a fully protonated ELP at a low pH (*17*). Here, sequence is not varied, but pH is used to vary charge density. The buffered solution contained 50 or 140 mM NaCl, which shields some of the long-range effects of charge.

2.2 Development of an equation for the binodal curve for dilute solutions

Mean-field (Flory-Huggins) theory can be used to construct a phase diagram for polymer solutions, which includes a binodal curve. The binodal phase boundary is the common tangent of the free energy of mixing at the equilibrium compositions of globule (Φ') and sediment (Φ'')

$$\left(\frac{\partial \Delta \overline{F}_{mix}}{\partial \phi}\right)_{\phi=\phi'} = \left(\frac{\partial \Delta \overline{F}_{mix}}{\partial \phi}\right)_{\phi=\phi''} \quad \textbf{(4)}$$

where $\Delta \bar{F}_{mix}$ is the free energy of mixing. The free energy of mixing is defined by the Flory Huggins equation for polymer solutions:

$$\Delta\overline{F}_{mix} = kT\left[\frac{\phi}{N}\ln\phi + (1-\phi)\ln(1-\phi) + \chi\phi(1-\phi)\right], \quad \textbf{(5)}$$

where k is the Boltzmann constant, N is the number of monomers in a polymer chain, Φ is volume fraction, and χ is the interaction parameter. The Flory Huggins equation models the competition between mixing entropy and non-specific solvent-polymer interactions. The first two terms promote mixing via combinatorial entropy, while the last term is energetic and describes the role of interactional enthalpy or entropy on mixing.

Mean-field theory is expected to fail in dilute solutions where the small overlap parameter makes concentration fluctuations large (*18*). Instead, dilute globules can be described as droplets whose surface tension ensures that they are roughly spherical. The relationship between sediment and globule concentrations can be described by equating the chemical potential of globule and sediment phases. Solving for surface tension,

$$\gamma \approx \frac{kT}{R_g^2}\ln\left(\frac{\phi''}{\phi'}\right) \quad \textbf{(6)}$$

where R_g is the radius of gyration of the globule. This is the surface tension along the phase boundary and relates the surface entropy of globules to the entropy difference between globules and sediment. More generally, surface tension is on the order of kT per thermal blob at the surface ξ_T^2:

$$\gamma \approx \frac{kT}{\xi_T^2} \approx \frac{kT}{b^8}v^2 \quad \textbf{(7)}$$

where v is excluded volume and c is a fitted parameter. b is the Kuhn length, the length of a monomer for a real freely-jointed polymer chain consisting of N segments. For length scales smaller than the Kuhn length, the polymer behaves more like a stiff rod. The Kuhn length of ELP was found to be 2.1 nm assuming an amino acid length of 0.365 nm (*19*).

For sizes larger than thermal blob size ξ_T, the size of the globule is related to the size ξ_T and length g_T of the thermal blob with the power law relationship,

$$R_g \approx \xi_T \left(\frac{N}{g_T} \right)^{1/3} \approx b \left(\frac{|v|}{b^3} \right)^{-1/3} N^{1/3}, \tag{8}$$

where 1/3 is the Flory scaling exponent for poor solvents. Combining the two equations for surface tension, we arrive at an excluded volume-concentration-length relationship:

$$\left(\frac{|v|}{b^3} \right)^{4/3} = \frac{1}{N^{2/3}} \ln \left(\frac{\phi''}{\phi'} \right). \tag{9}$$

Excluded volume is related to the θ temperature, an "intrinsic" temperature value. At θ temperature, chains have nearly ideal conformations at all concentrations. θ can also be described as the critical temperature for an infinitely long chain. For our naïve model, we can use an approximation of the temperature dependence of excluded volume in the neighborhood of θ temperature,

$$\frac{|v|}{b^3} \approx c_1 \frac{T - \theta}{T}. \tag{10}$$

In poor solvent, the temperature-concentration-length relationship becomes,

$$\left(c_1 \frac{T-\theta}{T}\right)^{4/3} = \frac{1}{N^{2/3}} \ln\left(\frac{\phi''}{\phi'}\right). \quad \textbf{(11)}$$

The final equation describes length and concentration effects on transition temperature, where sequence dependence is entirely described by a single parameter, θ. Here, we can see how measuring phase transition temperature can be a useful way to determine sequence contributions to excluded volume. This, in turn can be used to predict surface tension at the interface between protein polymer and water at physiologically relevant concentrations and temperature ranges, as they are generally in the vicinity of θ. Finally, we can approximate concentration of the sediment,

$$\phi'' \approx -c_2 \frac{v}{b^3} \approx c_2 \left(c_1 \frac{T-\theta}{T}\right). \quad \textbf{(12)}$$

2.3 Sequence dependence of the equation for an LCST binodal using experimental observations of a family of ELPs

We first fit the model to a large set of well-characterized phase-transition data of ELPs with varying alanine content, lengths, concentrations and transition temperatures in order to determine the relationship between θ and sequence content (*20*). We calculated volume fraction from mass concentration using a density of approximately 1 g/cm^3, and calculated N, number of Kuhn monomers by multiplying the number of amino acids by the ratio of Kuhn length to amino acid length, 2.1/0.365. To arrive at parameter estimations of c_1, c_2 and θ, we used nonlinear least squares fitting, minimizing error in Φ',

$$\ln(\phi') = \ln\left(c_2\left(c_1\frac{T-\theta}{T}\right)\right) - N^{2/3}\left(c_1\frac{T-\theta}{T}\right)^{4/3}. \quad \textbf{(13)}$$

This model was used to independently fit each previously characterized compositional libraries based on 2-3 peptide lengths (200 to 800 amino acids) and five molar concentrations (5 to 100 μM). Each composition is defined by their fractional alanine content in the guest residue (fA) (*20*). We used a subset of lengths in order to avoid extrapolated transition temperatures above 100°C, which are subject to a higher level of error.

We first noticed that an empirically-motivated modification of the excluded volume term by scaling it to the 0.6 power resulted in greatly improved fits, with average residual norms of log(Φ') of 0.54;

$$\frac{|v|}{b^3} \approx c_1\left(\frac{T-\theta}{T}\right)^{\nu}, \quad \textbf{(14)}$$

where $\upsilon = 0.6$. If we return to our equation for the temperature-dependence of surface tension,

$$\gamma \approx \frac{kT}{b^8}v^2 = \frac{kT}{b^2}\left(c_1\left(\frac{T-\theta}{T}\right)^{0.6}\right)^2 = \frac{kT}{b^2}c_1^2\left(\frac{T-\theta}{T}\right)^{1.2}, \quad \textbf{(15)}$$

we can see that this equation differs from our naïve expectation and is similar to the semi-empirical equation proposed by van der Waals, as well as Guggenheim and Katayama (*21*) for liquid surface tension,

$$\gamma \approx \gamma_0 \left(\frac{T_c - T}{T_c} \right)^n , \tag{16}$$

where T_c is the critical temperature, γ_0 is a liquid dependent parameter, and n = 11/9 for organic liquids. Ferguson found that for 14 common organic liquids n varied with a few percent only from 1.21, which is close to 11/9 (*22*). A more modern and theoretical basis for this empirical approach is based on renormalization group theory, which describes how fluids close to critical region exhibit large density fluctuations. The exponent n is associated with the correlation length of density fluctuations through the hyperscaling relation $n = (d - 1)\upsilon$ where $d = 3$ is the dimensionality of the system and $\upsilon = 0.629$ is the correlation length critical exponent, as predicted by the three-dimensional Ising model (*23*–*25*). That is, correlation length is the length scale of concentration fluctuations,

$$\xi \sim \left(\frac{T_c - T}{T_c} \right)^{-\nu} , \tag{17}$$

and surface tension of the globule is related to the length scale of concentration fluctuations ξ,

$$\gamma \sim \frac{1}{\xi^2} \sim \left(\frac{T_c - T}{T_c} \right)^{2\nu} , \tag{18}$$

where $2\upsilon = 2(0.629) = 1.258$, which is close to 11/9. Motivated by the form of this equation, we continue to use the correlation length critical exponent in our equations, as

$$\frac{|v|}{b^3} \approx \frac{b}{\xi_T} , \tag{19}$$

and,

$$\frac{|v|}{b^3} \approx c_1 \left(\frac{T-\theta}{T} \right)^{\nu} \approx c_1 \left(\frac{T-\theta}{T} \right)^{0.6} . \quad \textbf{(20)}$$

Among independent fits of compositional libraries, c_2 varied within a factor of 2 around the value of 1, so we fixed its value at 1. We verified that this resulted in physically relevant values of Φ" between 0.1 and 1, depending on T and θ. We also ensured that Φ" was always greater than critical $\Phi^* = 1/\sqrt{(N)}$. Φ_c was experimentally found to be around 0.15 volume fraction for 250 pentamers of VPGVG, which matches well with $1/\sqrt{(N)}$ without any prefactors (*13*). Thermal blob size, ξ_T, was also verified to be larger than Kuhn length b and smaller than the entire chain in ideal conformation, $b\sqrt{N}$, as,

$$\frac{\xi_T}{b} \approx \frac{b^3}{|v|} \approx \frac{1}{c_1} \left(\frac{T-\theta}{T} \right)^{-0.6} . \quad \textbf{(21)}$$

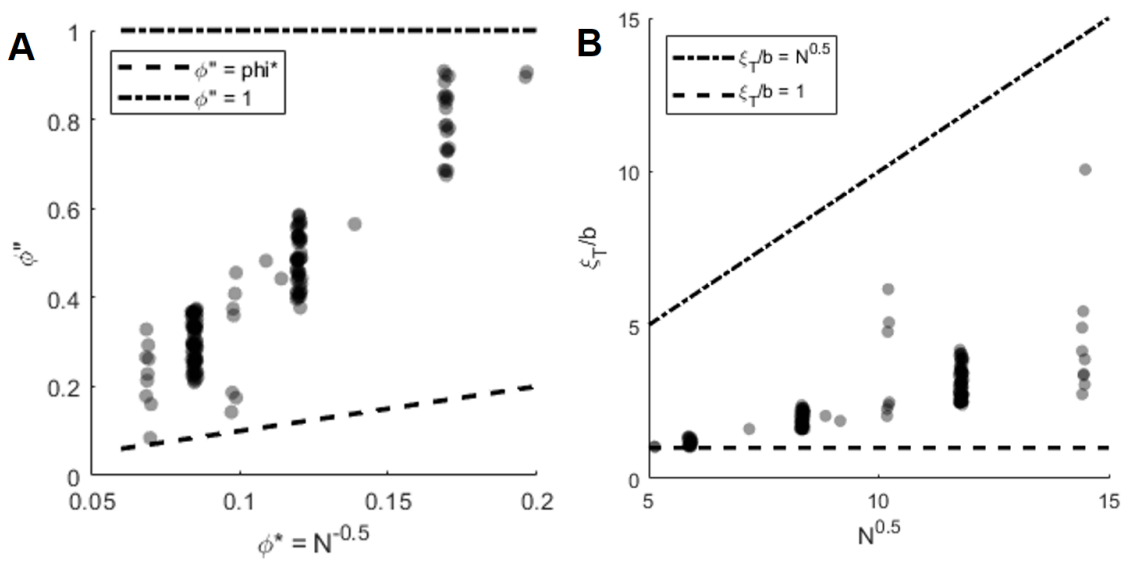

Figure 2: Verification of (A) Φ'' and (B) ξ_T/b. Both parameters are plotted against varying length, which indicate reasonable ranges of physical values. Dotted lines indicate defined boundaries of reasonable values.

Unexpectedly, however, c_1 of our model equation used to derive θ, also drifted with changing fraction of alanine. A double reciprocal plot of $1/c_1$ versus $1/\theta$ reveals a near linear relationship ($R^2 = 0.990$), resulting in the following equation,

$$\frac{1}{c_1} = -855.5\left(\frac{1}{\theta}\right) + 3.138\,. \tag{22}$$

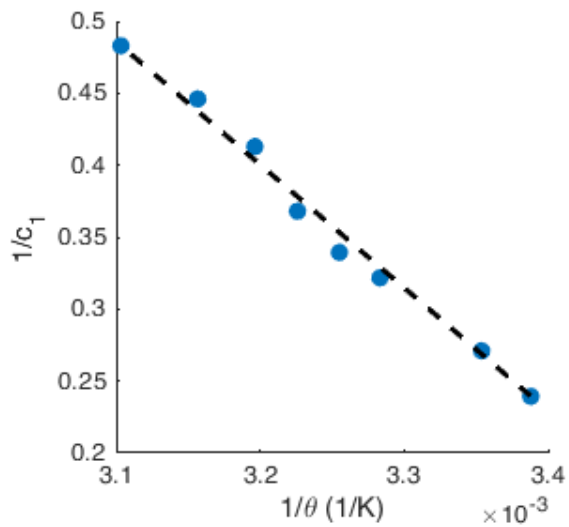

Figure 3: $1/c_1$ and $1/\theta$ are coupled and linearly related. Dotted line indicates a linear fit for fA ranging from 0 to 0.8.

Therefore, we reorganize the equation resulting in the following fitted unitless term for c_1 that explains its θ dependence,

$$c_1 = c\left(\frac{\theta}{\theta - \theta_0}\right) = \frac{1}{3.14}\left(\frac{\theta}{\theta - 273}\right), \tag{23}$$

where $\theta_0 = 273$ K refers to a fitted temperature reference point, and c is a fitted value of 3.14. The fitted constants θ_0 and c are remarkably close to the freezing point of PBS and π, respectively. Serendipitously, the fitted value for θ_0 to 5 significant figures is 272.63, and based on Blagdon's law,

$$T_f = 273.15 - iK_f m = 273.15 - 2 \cdot 1.86 \cdot 0.14 \text{ K} = 272.63 \text{ K}, \quad \textbf{(24)}$$

where i is the van't Hoff factor, K_f is the cryoscopic constant, and 0.14 is the molality of NaCl in PBS. Whether the closeness of these constants is coincidental or not can be an interesting subject for future investigation as its physical interpretation becomes understood.

We notice that our fitted equation of c_1 varies linearly with fraction of alanine ($R^2 = 0.998$), as

$$3.14 \cdot c_1 = \frac{\theta}{\theta - 272.63} = -6.757 f_{Ala} + 13.03, \quad \textbf{(25)}$$

as shown in Figure 4.

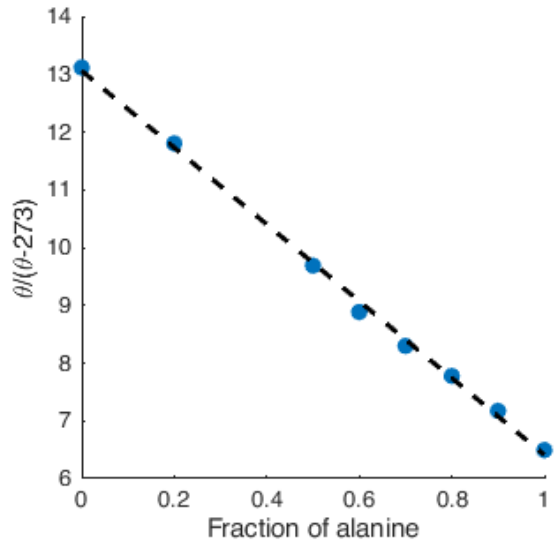

Figure 4: Fitted values for 3.14 c_1 versus fraction of alanine vary in a near linear manner with increasing substitution of valines to alanines. Dotted line indicates a linear fit.

We arrive at the following empirically-motivated modified excluded volume,

$$\frac{|v|}{b^3} \approx c\left(\frac{\theta}{\theta-\theta_0}\right)\left(\frac{T-\theta}{T}\right)^{\nu} = \frac{1}{3.14}\left(\frac{\theta}{\theta-273}\right)\left(\frac{T-\theta}{T}\right)^{0.6}, \tag{26}$$

and a corrected temperature-concentration-length relationship:

$$\frac{1}{3.14}\left(\frac{\theta}{\theta-273}\right)\left(\frac{T-\theta}{T}\right)^{0.6} = \left(\frac{1}{N^{2/3}}\ln\left(\frac{\phi''}{\phi'}\right)\right)^{3/4}. \tag{27}$$

Since Φ'' is temperature dependent,

$$\phi'' \approx -c_2\frac{v}{b^3} \approx \frac{1}{3.14}\left(\frac{\theta}{\theta-273}\right)\left(\frac{T-\theta}{T}\right)^{0.6}, \tag{28}$$

(27) can be rewritten as an explicit fitting equation for Φ',

$$\phi' = \phi'' \exp\left[-N^{2/3} \left(\frac{1}{3.14} \left(\frac{\theta}{\theta - 273} \right) \left(\frac{T-\theta}{T} \right)^{0.6} \right)^{4/3} \right]. \quad \textbf{(29)}$$

It is worth noting that the contribution of the sediment entropy is often negligible as the sediment volume fraction usually larger than that of globules in the supernatant (*26*).

Finally, we note the temperature dependence of the following physical quantities using the modified excluded volume: for radius of gyration,

$$R_g \approx b \left(\frac{|v|}{b^3} \right)^{-1/3} N^{1/3} \approx \frac{bN^{1/3}}{\left(\frac{1}{3.14} \frac{\theta}{\theta - 273} \right)^{1/3} \left(\frac{T-\theta}{T} \right)^{0.2}}, \quad \textbf{(30)}$$

and surface tension,

$$\gamma \approx \frac{kT}{R_g^2} \ln\left(\frac{\phi''}{\phi'} \right) \approx \frac{kT}{b^2 N^{2/3}} \ln\left(\frac{\phi''}{\phi'} \right) \left(\frac{1}{3.14} \frac{\theta}{\theta - 273} \right)^{2/3} \left(\frac{T-\theta}{T} \right)^{0.4}. \quad \textbf{(31)}$$

We verified that radius of gyration varied between 5 and 25 nm, and surface tension varied between 0.01kT and 0.2kT, ranges which are within reasonable physical values.

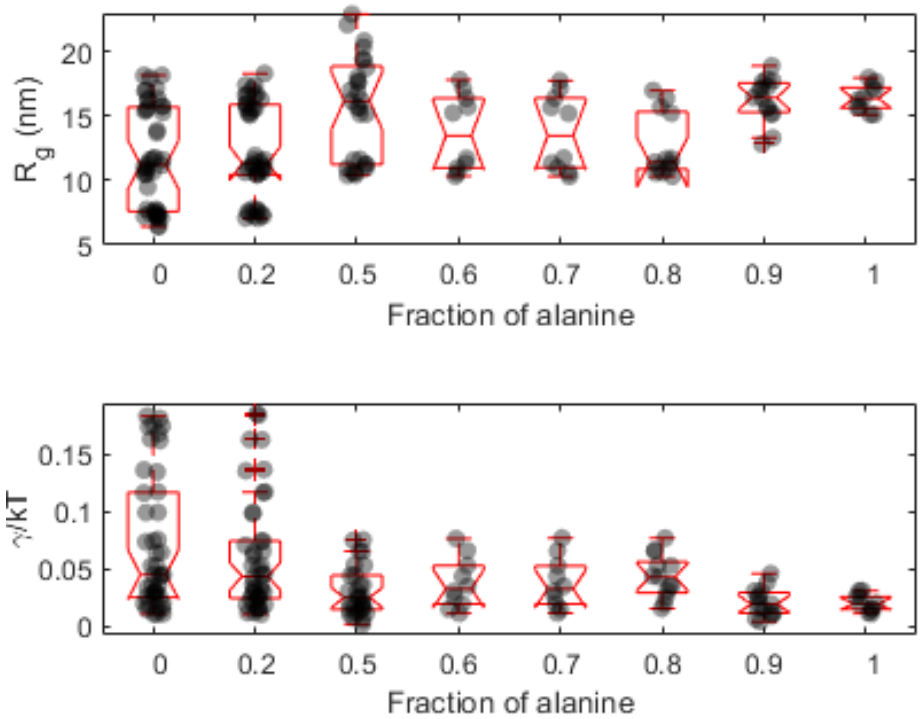

Figure 5: Verification of R_g and γ/kT. Box plots of both parameters for independent fits of the model across varying fraction of alanine indicate reasonable ranges of physical values.

Interestingly, modification of the excluded volume introduces a new reference point θ_0 = 272.63 K whose temperature value is a lower limit for θ which is remarkably close to freezing point of the solvent. The ratio between θ and 272.63 K value appears to modulate the miscibility gap in the temperature-concentration phase diagram, where lower values of θ have wider binodal curves as it approaches θ = 272.63 K until it becomes flat, and may suggest a connection between LCST and freezing. Figure 6 shows the collapse of multiple length data sets for each sequence with varying fraction of alanine. The fitted models are separate lines in this space. As θ is lowered, the slope decreases, which is indicative of broad binodals in the phase diagram.

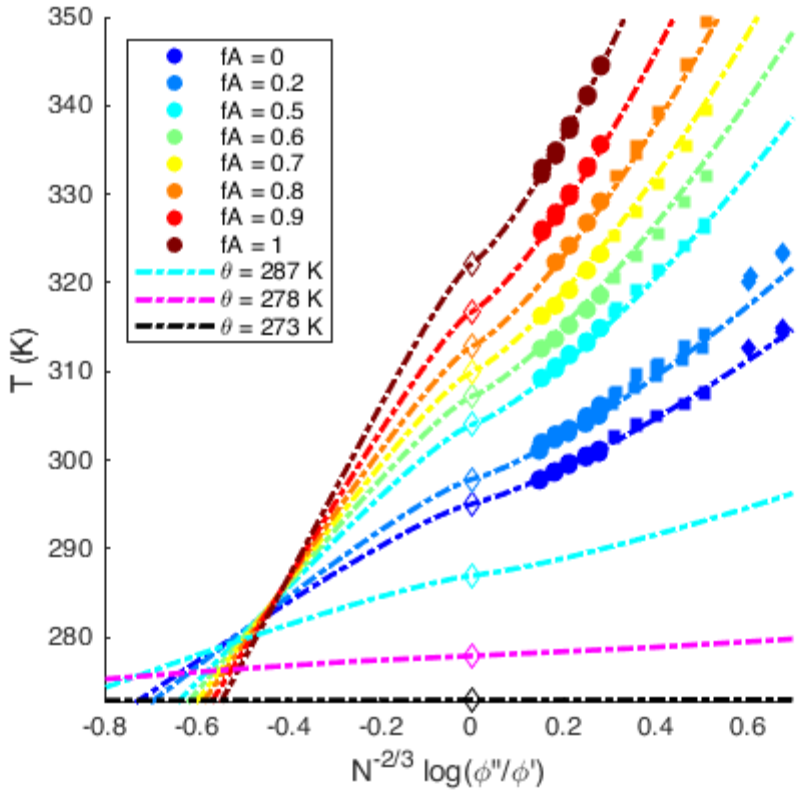

Figure 6: Temperature dependence of the length-collapsed miscibility gap. The plot shows the collapse of varying length and concentration data for independent fits for a family of sequences. Dotted lines are fits of data or theoretical constructs. Circles represent lengths of 160 amino acid pentamers, squares represent lengths of 80 pentamers, and filled diamonds represent lengths of 40 pentamers. Open diamonds show the θ points.

Finally, plotting the left and right hand sides of the following reformatted equation,

$$\left(1-\frac{\theta}{T}\right)^{0.6} = \frac{1}{\sqrt{N}}\ln\left(\frac{\phi''}{\phi'}\right)^{3/4} 3.14\left(1-\frac{273}{\theta}\right), \tag{32}$$

helps visualize the collapse of data. For smaller lengths (40 pentamers), the data starts to diverges from a linear relationship, which suggests a potential crossover across the

critical region, as determined by the Ginzburg criterion. This divergence can potentially be corrected with a crossover function (*24*).

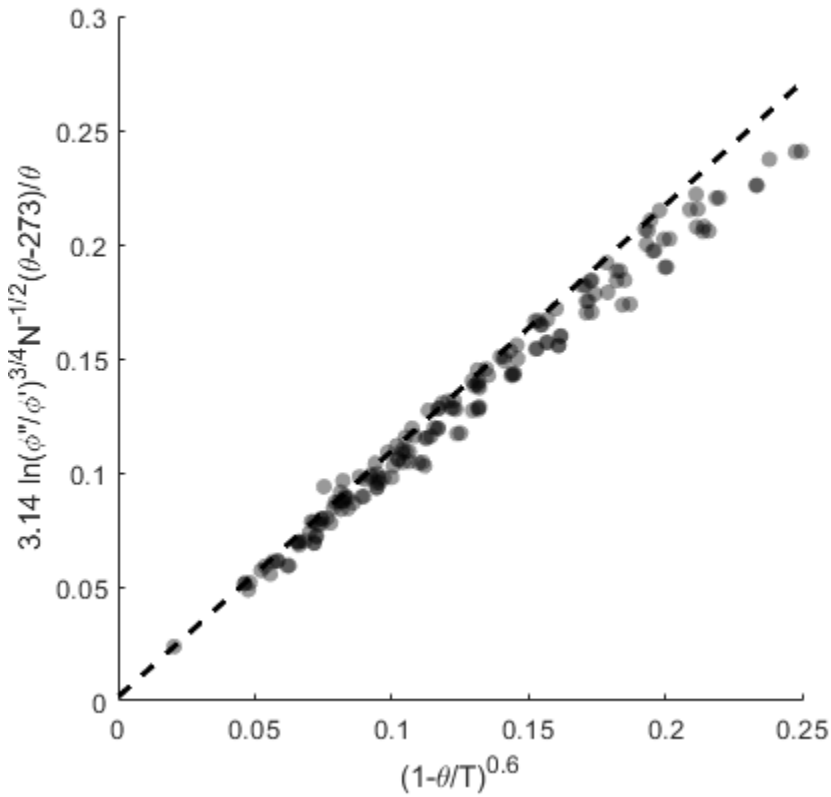

Figure 7: Visualization of the collapse of data. Plotting the temperature dependent term versus the rest of the equation collapses data along a line. Dotted line indicates a linear fit of data with temperatures close to θ.

2.4 Observations on θ and θ_0 for a more general set of IDPPs

With a general model equation, we can now apply to a broader range of sequences with more limited concentration and length variation, but wider variation in sequence content. We allowed both θ and θ_0 to be fitting parameters instead of fixing θ_0 = 272.63 K to examine the effects of a wider range of sequences on the model. These data

were taken primarily from IDPP work by Quiroz and Chilkoti (*6*) and others (*27–39*, *39*). It is worth noting that, in this set of data, there is a predictable bias of amino acid sequence content towards aliphatic amino acids, as their hydropathies (*37*) are within a relevant range. This is due to the requirement of phase transition within the range of physiological temperatures and bounded in concentration by the yield-limitations of protein purification.

By fitting families of data sets independently based on sequences with varying length and/or concentration, fitted values of θ and θ_0 can be obtained for each repeated IDPP motif.

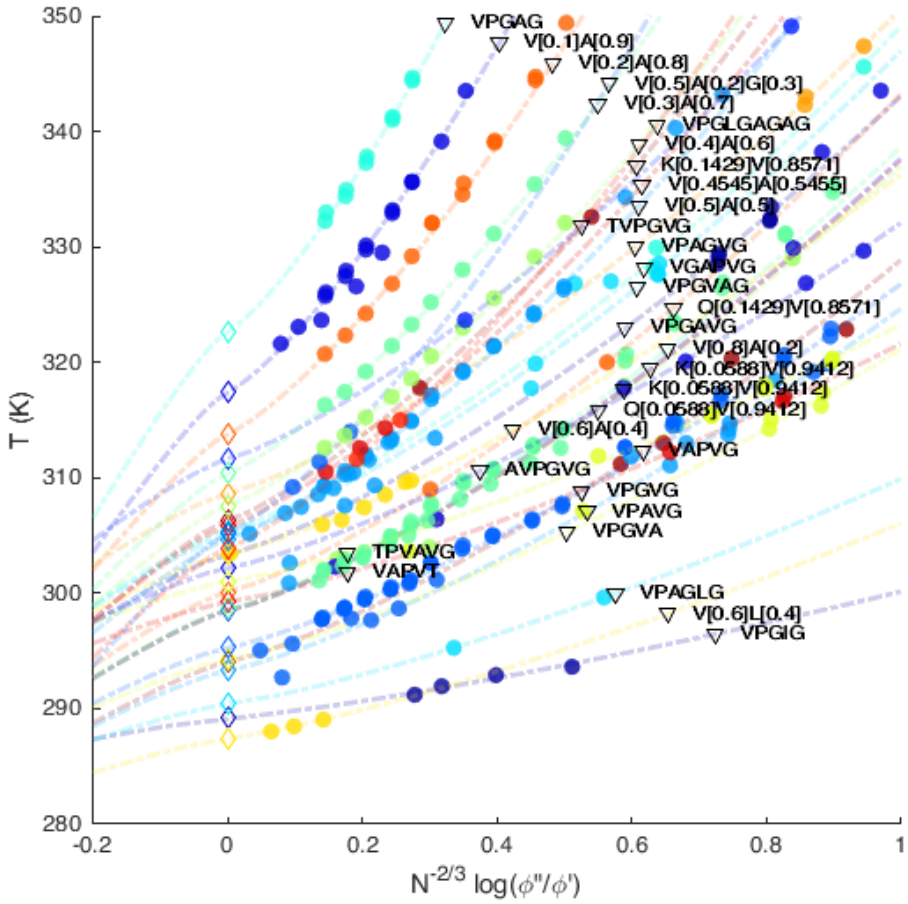

Figure 8: Linearization of the equation for *T* for independent fits for a variety of different sequence patterns. Labels consisting of letters indicate the repeated amino acid motif patterns, whereas labels consisting of letters and numbers, e.g. X[f_1]Z[f_2], describe the guest residues of ELPs where X, Z, … indicate amino acids and f_1, f_2, … indicate their respective fractions. Filled circles indicate data points, dotted lines indicate fits, open diamonds indicate θ temperatures, and open squares indicate θ_0 temperatures.

We found that equation (23) had was related linearly to the contributions of each amino acids for this set of sequences. We performed a multivariate regression, resulting in the following fit equation:

$$c_{1,\text{linear}} = \frac{\theta}{\theta - \theta_0} = 1.20 f_{Ala} - 47.7 f_{Gln} - 3.96 f_{Gly} + 88.8 f_{Ile} + 68.9 f_{Leu} - 103.1 f_{Lys} + 1.13 f_{Pro} + 8.33 f_{Thr} + 36.8 f_{Val}, \quad \textbf{(33)}$$

where f_{Xaa} indicates the fraction of Xaa residue in an IDPP sequence. Note that the predictive strength of coefficients for nonaliphatic amino acids is weak due to the limited variation in the data set. However, c_1 and $c_{1,\text{linear}}$ were well-correlated ($R^2 = 0.94$), indicating that the contributions of each amino acid of this family of aliphatic amino acids to this ratio may be additive.

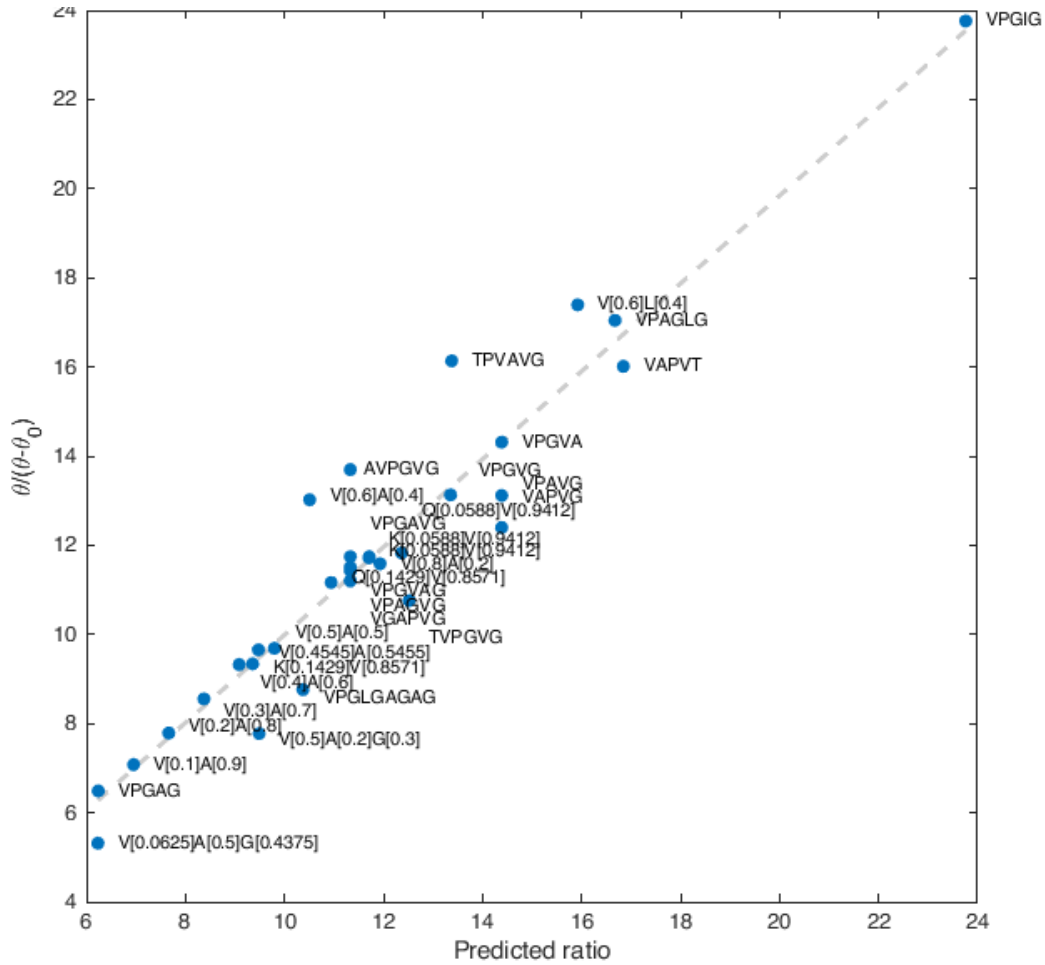

Figure 9: Linear correlation between fitted c_1 and $c_{1,\text{linear}}$ from a multivariate-linear model. Dotted line indicates linear regression. Labels indicate the sequence pattern for each independently fit data set.

A plot of θ_0 versus $\theta/(\theta-\theta_0)$ indicates lack of correlation ($R^2 = 0.17$), signifying that θ_0 is less likely simply explained by additive contributions of amino acids. Some of the variation in θ_0 may be attributed to extrapolation error due to the fact that the majority of data was measured on low concentrations (~25 μM) and lengths (~40 pentamers). However, upon visual inspection, we can see a clustering of θ_0 around 273 K.

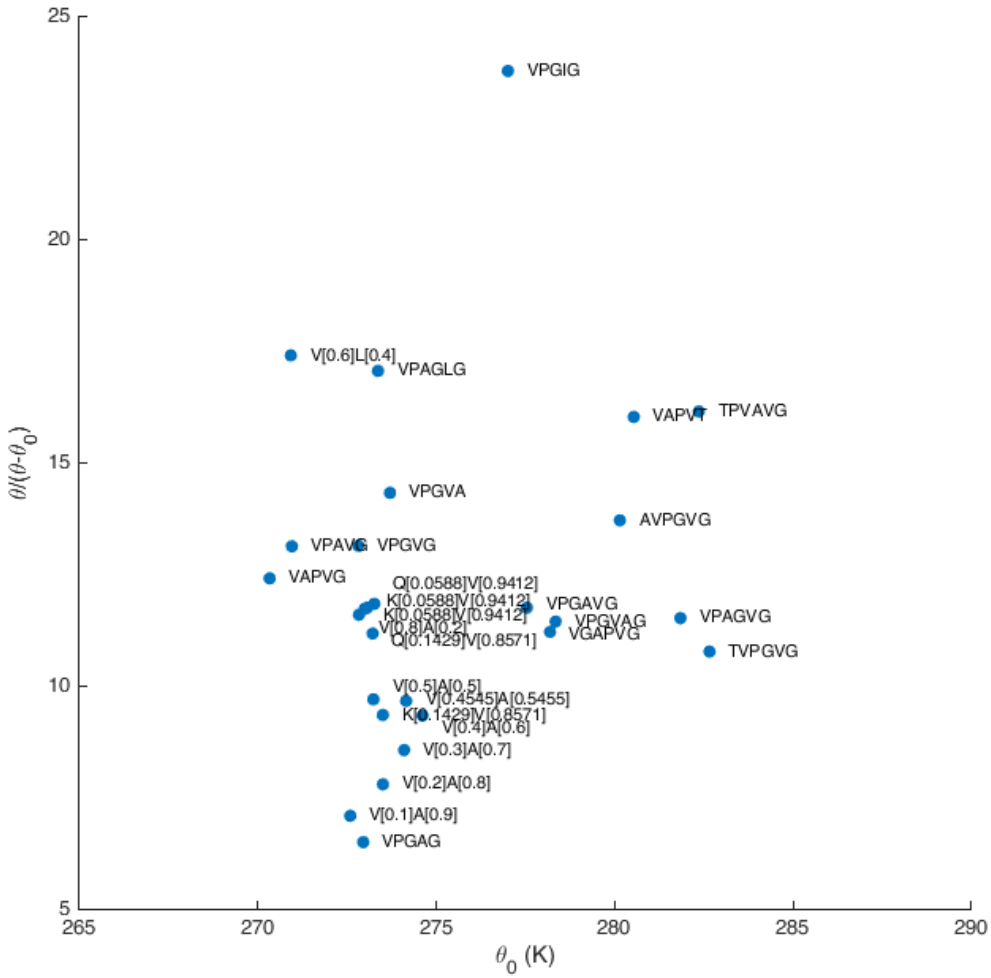

Figure 10: Lack of correlation between θ and θ_0. Labels indicate the sequence pattern for each independently fit data set. There is a dense clustering of data around θ_0 = 273 K.

We conjecture that deviations from the clustering of sequences around θ_0 = 273 K may result from ordering and distribution of amino acid, which in turn relate to chain-chain interactions and/or hydration. For example, some of the sequences with varying sequence order and varying θ_0 have hysteretic transition behavior which may be related to changes in level of hydration. Also, recall that sequences of the form V[x]A[y] represent various fractions and distributions of 'V' blocks and 'A' blocks of the form

VPGVG and VPGAG, respectively, which vary in block-distribution e.g. AAVAVAAVAA for V[0.3]A[0.7]. We propose that slight variation in θ_0 for this group of sequences may be attributed to changes in chain-chain interactions resulting from this effect.

We also consider sequences containing histidine residues, V[0.2]H[0.8] and V[0.2]H[0.4]G[0.2]A[0.2]. Upon lowering of pH from 8 to 6.4, $\theta/(\theta - \theta_0)$ remained fairly constant, while both θ and θ_0 values increased, resulting in a near vertical trajectory in the $\theta/(\theta - \theta_0) - \theta_0$ space (Figure 12a). Variation in θ_0 appears to be directly related to charge effects, as the variation in θ_0 with pH follows a fit based on a linear transformation of the Henderson Hasselbach equation,

$$\theta_0 = A + B\left(1 - f_+\right) = A + \frac{B}{1 + 10^{pKa - pH}}, \quad \textbf{(34)}$$

where A refers to θ_0 at protonated state, B refers to the difference between θ_0 at protonated and deprotonated state, f_+ refers to the fraction of Histidine residues that are protonated, and pKa was obtained through fits, at a value of 6.75 (*39*). A nonlinear least squares fitting resulted in A values of 276 K and 275 K and B values of -34.8 and -27.5 for V[0.2]H[0.8] and V[0.2]H[0.4]G[0.2]A[0.2] respectively. Note that due to the presence of ~140 mM NaCl in the PBS solution, some of the long-range effects of charge are electrically screened.

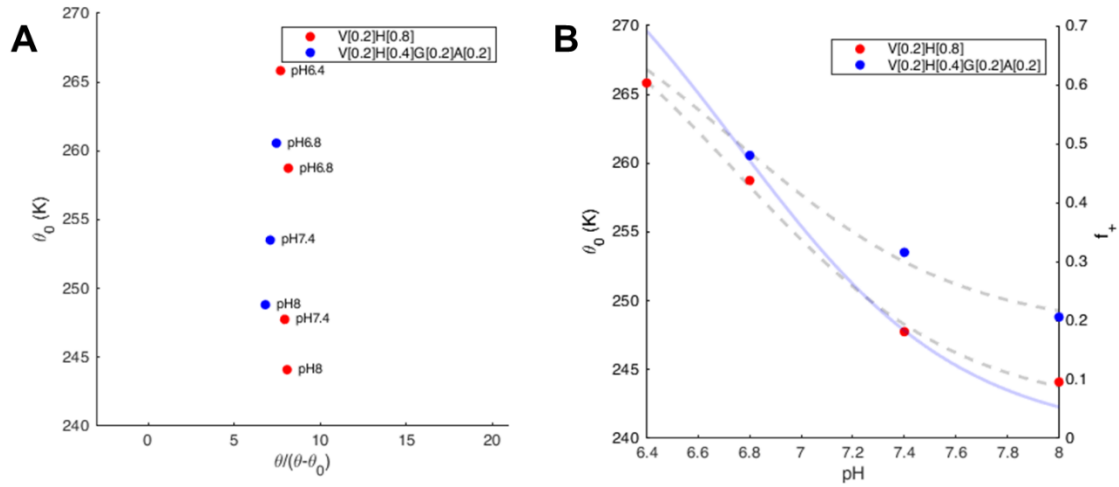

Figure 11: Trajectories in the $\theta/(\theta - \theta_0) - \theta_0$ space for increasing pH (A) and the nonlinear effects of pH on θ_0 (B) for histidine-containing sequences. Note the flipped axis in (A) compared to Figure 10 to prevent overlapping labels, and that the axis scale ratio is preserved. Dotted lines in (B) indicate fits based on the Henderson Hasselbach equation. The solid line indicates the estimates of the protonated fraction of Histidines f_+ for both constructs.

To observe the effects of sodium chloride in the solvent, we considered the ELP sequence, V[0.2]A[0.8], as it was hydrophilic enough that the addition of sodium chloride would not depress phase transition temperatures below observable temperature ranges. Again, the addition of salt results in a near vertical trajectory in the $\theta/(\theta - \theta_0) - \theta_0$ space, with a minimal changes in $\theta/(\theta - \theta_0)$. However, θ_0 varied almost linearly with sodium chloride concentration with the following fit for concentrations between 0.75 M and 2 M ($R^2 = 0.998$),

$$\theta_0 = 276 - 10.23[\mathrm{NaCl}]. \quad \textbf{(35)}$$

The fit also appears to overestimate temperatures for concentrations of sodium chloride lower than 0.75 M.

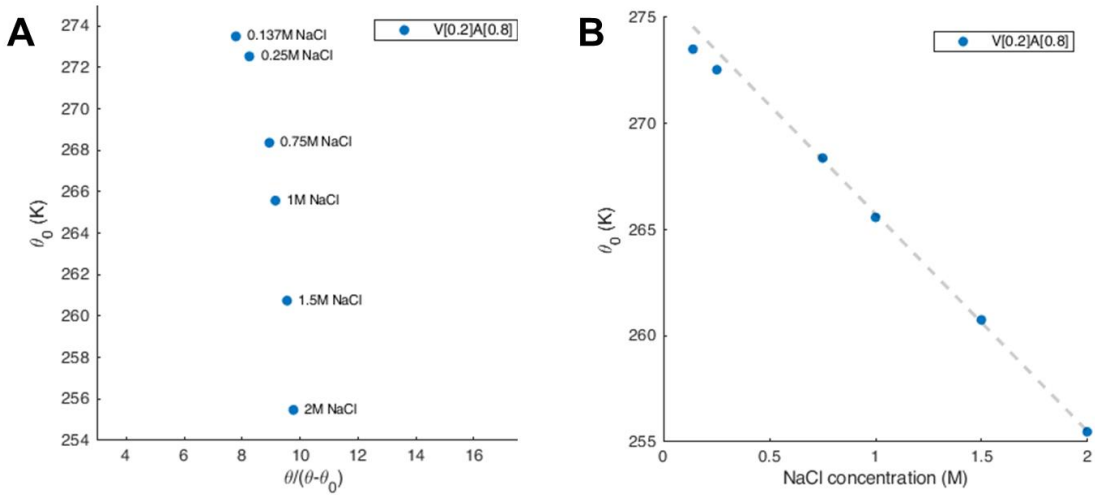

Figure 12: The trajectory in $\theta/(\theta - \theta_0) - \theta_0$ space for increasing salt (A) and the linear effects of NaCl concentration on θ_0 (B) for V[0.2]A[0.8]. Note the flipped axis in (A) compared to Figure 10 to prevent overlapping labels, and that the axis scale ratio is preserved. The dotted line in (B) indicates a linear fit.

Based on our initial observations of the trajectories in the $\theta/(\theta - \theta_0) - \theta_0$ space, we conjecture that θ_0 is related to chain-chain interactions, and also entertain a potential relationship with freezing. The trajectories of changing variables such as sodium chloride concentration and charge appear nearly orthogonal to the cluster of elastin related motifs around $\theta_0 = 273$ K, and may be informative of the nature of θ_0. $\theta/(\theta - \theta_0)$ is potentially correlated with hydrophobicity, based on its role as a multiplicative term in the temperature dependence of excluded volume and surface tension, and based the observation that aliphatic amino acid contributions are additive. While a complete thermodynamic explanation of θ_0 and θ is beyond the scope of this work, it follows that additional studies on the role of tunable variables including charge, UCST, and salt may further improve our understanding of θ_0. In general, the $\theta/(\theta - \theta_0) - \theta_0$ space may be a

useful diagram for revealing trajectories with respect to the change of various variables including charge, sequence and salt concentration.

3. Codon scrambling: an enabling technology

Our engineering-focused goal is to develop generalized materials design rules, by combining high throughput experiments and physics-based models to accelerate the discovery and development of a class of thermally responsive polypeptide materials by a systematic exploration of sequence space. In addition, with the proposed model of Chapter 2, we can develop technologies to further validate it and understand the thermodynamic meanings of its "intrinsic", sequence-dependent parameters: θ and θ_0.

We would like to identify the sequence determinants of thermally responsive behaviour (charge, hydrophobicity, etc.). This is driven by scalable tools developed for directed evolution and rational design of peptide polymers. With a combination of two strategies (Chapter 3 and Chapter 4), we intend to continually improve the rate of discovery and characterization of new stimuli responsive peptide polymers.

Our first strategy, whose formulation is discussed in this chapter, serves to design novel and diverse sequences that can be easily commercially synthesized in medium throughputs. The codon-scrambling algorithm enables the PCR-based gene synthesis of repetitive proteins by exploiting the codon redundancy of amino acids and finding the least-repetitive synonymous gene sequence. The following chapter contains material which borrows heavily from published work (*40*).

3.1 Introduction

The production scale of repetitive genes continues to fall behind that of non-repetitive genes because they cannot leverage the recent technological advances in next-generation, automated gene synthesis. In the past 15 years, these advances, driven by advancements in synthetic biology, have enabled the market cost of commercial gene synthesis to drop by over 100-fold (*41*). Prices continue to drop by a factor of 1.5 per year, in a manner akin to Moore's law (*42*). By 2005, 39 gene synthesis vendors existed in the world, and the number has increased since then. In the past couple of years, more companies have begun to emerge to take on the challenge of possibly adapting new synthesis technologies to the market. Gen9 Inc. and Twist Biosciences Inc. have developed scalable gene synthesis technology platforms based on the assembly of DNA pools from microarray chips (*43*, *44*), which enables the fabrication of high throughput, low-cost 96-well microplates of ~1kb genes.

Therefore, the development of PCR-based repetitive gene synthesis is critical to efficiently synthesize repetitive genes with high throughput (*45*). PCR-based assembly, otherwise known as polymerase cycling assembly (PCA), remains one of the most cost-effective gene assembly methods (*46*, *47*). In a two-step procedure, partially overlapping oligonucleotides are designed span the whole sequence. A PCR reaction is carried out so that overlapping oligonucleotides anneal and extend. The resulting double stranded construct can serve as the template for PCR amplification reaction.

In order to synthesize repetitive proteins via commercially compatible gene assembly approaches, we have developed a codon scrambling algorithm that exploits

the codon redundancy of amino acids and identifies the least repetitive synonymous coding sequence for any protein sequence. Given the exponentially large number of codon variants, the search for the least repetitive sequence in an immense sequence space is a computational challenge. Although this discrete optimization problem belongs to the non-deterministic polynomial-time (NP)-hard complexity class, we found that developing a modestly sized problem formulation was crucial to solving it without resorting to metaheuristic algorithms, which are approximate and usually non-deterministic (*48*).

The objective of the codon scrambling problem is to minimize potential cross-hybridization events, or off-target interactions, of a repetitive coding sequence during polymerase reactions. In this problem, a polymerase reaction is conceptualized as a set of local cross-hybridization events between repeated subsequences. The tendency of duplex formation is estimated by the exponential-transformed duplexing energy ΔG given by the Boltzmann weight $\exp(-\Delta G/RT)$, where T is the system's absolute temperature and R is the gas constant. We can then quantify tradeoffs between sequences with a scalarized objective function, defined by the linear combination of all subsequence repeats weighted by their Boltzmann weights.

We utilize the *codon De Bruijn subgraph*, an adaptation of the De Bruijn subgraph (*49*), to sparsely represent all feasible nucleotide sequences that encode a protein of interest. A k-mer De Bruijn subgraph of a 64 codon alphabet $\circledcirc_{CODON}$ is a directed graph in which each vertex represents a substring consisting of k codons, or $3k$ nucleotides. Each arc is a substring consisting of $k + 1$ codons, and connects its prefix and suffix vertices;

for example, arcs between vertices corresponding to dicodons can be represented by tricodons, as shown by the overlapping nucleotide sequences below:

$$GGCGTA \xrightarrow{GGCGTA\boldsymbol{CCT}} GTA\boldsymbol{CCT}\ , \tag{36}$$

where $k = 2$. In this sense, a path traversal across multiple arcs represents a *sliding window*, where a window of $k + 1$ codons incrementally advances across a sequence. We can then construct a graph such that feasible paths code the protein of interest, provided that the graph is defined by the order and composition of the amino acid sequence (Figure 13). The codon De Bruijn subgraph is a relaxed version of the De Bruijn subgraph which allows for substrings that appear more than once if they code for different parts of the amino acid sequence. Arcs can easily be deleted from the subgraph if they contain forbidden sequences such as homopolymeric stretches, strong secondary structures, internal ribosome binding sites (iRBS) (*50*), or restriction enzyme recognition sites of interest.

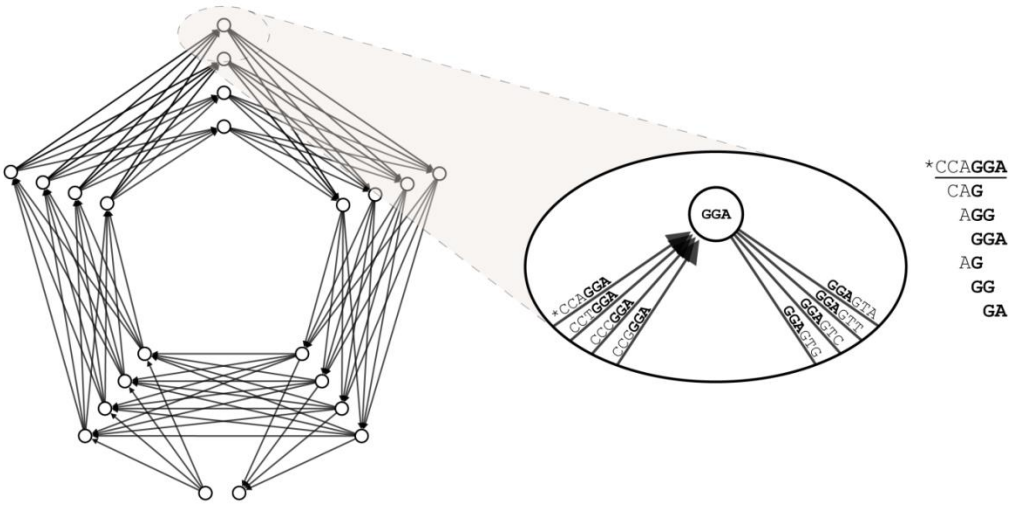

Figure 13: Example diagram of a multipartite *n*-mer codon De Bruijn subgraph with a window length *k* = 1 and a motif length *r* = 5 (GVGVP motif). Notice that the intermediate arcs can be clearly separated into *r* partitions. Here, there are 4 possible codon alternatives for each amino acid. Each vertex has 8 incident arcs, and each arc can be decomposed into codon-level suffixes which are then used to calculate the objective function. From (*40*).

In this graph, each arc is decomposed into a set of subsequences which are used to calculate the arc's (non-additive) contribution to the objective function. Therefore, the fixed-length path with the minimum objective defines the least repetitive codon variant. We have found that the discrete optimization on the set of all feasible paths is analogous to the classical traveling salesman problem (*51*). This is due to our utilization of *subtour elimination constraints* (*51*), used to eliminate disjoint cycles from the set of feasible solutions, as well as the requirement of simple, fixed-length paths. We provide a detailed description of the mixed integer linear program (MILP) formulation in the following section, in which we utilize the graph's network topology to reduce the size and difficulty of the problem. In its implementation, we use MATLAB (Mathworks) code to convert repetitive protein sequences into MILP problem instances, which are then

optimized using the IBM ILOG CPLEX Optimization Studio 12.5.1, a modern state-of-the-art MILP solver (http://chilkotilab.pratt.duke.edu/codon-scrambler).

3.2 Formulation of the codon scrambling problem

Informally, the codon scrambling problem asks for the codon variant which has the minimum weighted sum of nucleotide subsequence repeats. For a given codon variant, each subsequence m is associated with its Boltzmann weight c_m and its number of repeats $\varphi(m)$. As defined earlier, the Boltzmann weight, $c_m = e^{-\Delta G_m / RT}$. For a subsequence that occurs at least once, the number of direct and inverted repeats is,

$$\varphi(m) = \begin{cases} \nu(m) - 1, & \nu(m) \geq 1 \\ 0, & \text{otherwise,} \end{cases} \tag{37}$$

where nucleotide multiplicity $\nu(m)$ is the total number of occurrences of a nucleotide sequence and its reverse complement. Let each codon variant be represented by its multiset of subsequences M. Then, the minimum weighted sum of subsequence repeats is:

$$\sum_{m \in M} c_m \varphi(m), \tag{38}$$

For each codon variant, we can naively determine M by traversing the entire codon variant generating nucleotide subsequences of length $2 \ldots n$. However, given the exponentially large number of codon variants, we instead determine M using codon De Bruijn subgraphs. As defined earlier, each arc of the codon De Bruijn graph represents a $3(k + 1)$-mer subsequence, and each feasible path represents a codon variant. Consider that M is calculated from each feasible path. For each constituent arc of the feasible path,

$ij \in A'$, we construct the multiset S_{ij}: we decompose each 3(k + 1)-mer into *codon-level* suffixes, shorter subsequences that end with one of the last three nucleotides. In this way, a path traversal combines the codon-level suffixes into the aggregate multiset sum such that $M = \uplus_{ij \in A'} S_{ij}$ and the objective **(38)** is calculated.

Table 1 is the MILP formulation of the codon scrambling problem. Briefly, x_{ij} is equal to 1 if and only if arc $ij \in A$ is in the optimal path. Objective (a) and constraints (b and j) define $\varphi(m)$ using a binary slack variable, z_m. Constraint (c) ensures that the longest subsequences do not repeat. The remaining constraints are path constraints: constraints (d and i) impose that arcs are connected and distinct, constraint (e) imposes the path length, and constraints (f –h) impose that the path is acyclic.

Table 1: MILP formulation

Sets

V	set of vertices
A	set of arcs
$\mathcal{U}$	universe of substrings of length $\leq 3k$
$S \subseteq A$	feedback arc set
$K \subseteq \mathcal{U}$	set of substrings of length $< 3k$

Variables

x_{ij}	$ij \in A$	binary flow through arc ij
z_m	$m \in K$	binary slack variable for substring m
u_i	$i \in V \setminus \{s,t\}$	vertex potential at vertex i

Parameters

p		number of motif repetitions
ΔG_m	$m \in K$	duplexing Gibbs energy for substring m
w_{ij}^m	$m \in \mathcal{U}, ij \in A$	occurrences of substring m in arc ij
$b_i = \begin{cases} -1, \\ 1, \\ 0, \end{cases}$	$i = s$ $i = t$ $\forall i \in V \setminus \{s,t\}$	flow balance between source s and sink t
$c_{ij} = \sum_{m \in K} w_{ij}^m e^{-\Delta G_m / RT}$	$ij \in A$	pre-calculated terms in the piecewise objective
$d_m = e^{-\Delta G_m / RT}$	$m \in K$	

Objective

$$\min \quad \sum_{ij \in A} c_{ij} x_{ij} - \sum_{m \in K} d_m z_m \tag{a}$$

Constraints

$$z_m - \sum_{ij \in A} w_{ij}^m x_{ij} \leq 0 \qquad \forall m \in K \tag{b}$$

$$\sum_{ij \in A} w_{ij}^m x_{ij} \leq 1 \qquad \forall m \in \mathcal{U} \setminus K \tag{c}$$

$$\sum_{j \in V} x_{ji} - \sum_{j \in V} x_{ij} = b_i \qquad \forall i \in V \tag{d}$$

$$\sum_{ij \in S} x_{ij} = p - 1 \tag{e}$$

$$u_i - u_j + p x_{ij} \leq p - 1 \qquad \forall ij \in S \tag{f}$$

$$u_i - u_j + (p-1) x_{ij} \leq p - 1 \qquad \forall ij \in A \setminus S, i \neq s, j \neq t \tag{g}$$

$$0 \leq u_i \leq p - 1 \qquad \forall i \in V \setminus \{s,t\} \tag{h}$$

$$x_{ij} \in \mathbb{Z}_2 \qquad \forall ij \in A \tag{i}$$

$$z_m \in \mathbb{Z}_2 \qquad \forall m \in K \tag{j}$$

3.3 Development of the codon scrambling problem and its MILP formulation

In the CODON SCRAMBLING problem, we are formally given a directed graph $G = (V, A)$ with source vertex $s \in V$ and sink vertex $t \in V$, a multiset of nucleotide subsequences U, a set of positive real weights $C = \{c_m : m \in \mathcal{U}\}$, and a family S of multisets of U where $\mathcal{S} = \{S_{ij} \subseteq \mathcal{U} : ij \in A\}$. The problem asks for a simple st-path with subgraph $P = (V', A')$ and n intermediate vertices such that the objective $\sum_{m \in M} c_m \varphi(m)$, (38) is minimized where $M = \uplus_{ij \in A'} S_{ij}$. When the CODON SCRAMBLING problem is unweighted and unconstrained by path constraints, it asks the MINIMUM REPEAT MULTIPLICITY problem: given a multiset U and a family S of multisets of U, select a subfamily n multisets $\mathcal{E} \subseteq \mathcal{S}$ such that the sum over repeat multiplicities of elements in the multiset sum is minimized.

First, we prove that MINIMUM REPEAT MULTIPLICITY is NP-hard by a reduction from k-SPARSEST SUBGRAPH. The k- SPARSEST SUBGRAPH problem and its weighted version, the k-LIGHTEST SUBGRAPH problem, are NP-hard (*52*). It asks: given a simple graph $G = (V, E)$ and $k \leq |V|$, select k vertices that minimizes the number of edges of the induced subgraph.

Lemma 1. MINIMUM REPEAT MULTIPLICITY *is* NP-*hard*.

Proof. We show that k-SPARSEST SUBGRAPH $\leq_P$ MINIMUM REPEAT MULTIPLICITY. In our construction, let $\mathcal{U} = E$, $n = k$ and $\mathcal{S} = \{S_i \subseteq E : i \in V\}$ where S_i is the set of edges incident to vertex i. Suppose we have an induced subgraph $G[V'] = (V', E')$ where

$|V'| = k$ and $E' = \{uv \in E : u, v \in V'\}$. Let the subfamily $\mathcal{E} = \{S_i \subseteq E : i \in V'\}$. Then $|\mathcal{E}| = n$. Because G is a simple graph, uv is contained once each in only S_u and S_v for all $uv \in E'$. It follows that $\nu(uv) = 2$ and $\varphi(uv) = 1$ for all $uv \in \uplus_{S_i \in \mathcal{C}} S_i$. Let $M = \uplus_{S_i \in \mathcal{C}} S_i$. Then, the sum of repeat multiplicities, $\sum_{uv \in M} \varphi(uv) = |M| / 2 = |E'|$. Thus, we can compute the sparsest subgraph by computing the minimum sum over repeat multiplicities. □

MINIMUM REPEAT MULTIPLICITY can be formulated as an integer program. We define the binary slack variable $z_e \in \mathbb{Z}_2$ such that $\nu(m) = \varphi(m) - z_m$ for all $m \in \mathcal{U}$. Namely, $z_m = 1$ only when $\nu(m) \geq 1$ according to **(37)**. We also define the binary variable x_S such that it is equal to 1 if and only if $\mathcal{S} \in \mathcal{E}$. The parameter w_S^m is the integer multiplicity of element $m \in \mathcal{U}$ in multiset $S \in \mathcal{S}$; then, $\sum_{S \in \mathcal{S}} w_S^m x_S = \nu(m)$ and resulting integer program follows:

$$\begin{aligned} \min \sum_{S \in \mathcal{S}} c_S x_S - \sum_{m \in \mathcal{U}} z_m & \\ z_m - \sum_{S \in \mathcal{S}} w_S^m x_S \leq 0 & \qquad \forall m \in \mathcal{U}, \\ \sum_{S \in \mathcal{S}} x_S = n & \\ z_m \in \mathbb{Z}_2 & \qquad \forall m \in \mathcal{U}, \\ x_S \in \mathbb{Z}_2 & \qquad \forall S \in \mathcal{S}. \end{aligned} \tag{39}$$

where, $c_S = \sum_{m \in \mathcal{U}} w_S^m$ and is easily pre-calculated for all $S \in \mathcal{S}$.

Theorem 1. CODON SCRAMBLING *is* NP-*hard*.

Proof. We show that MINIMUM REPEAT MULTIPLICITY $\leq_P$ CODON SCRAMBLING. We have a universe $\mathcal{U}$ and a family $\mathcal{S}$ of multisets of $\mathcal{U}$. Let $G = (V, A)$ be a complete digraph

such that $V = \{s,t,i : S_i \in \mathcal{S}\}$ and $A = \{ij : S_i \in \mathcal{S} \cup \{S_s, S_t\}, S_j \in \mathcal{S} \cup \{S_s, S_t\}\}$ where $S_s = \varnothing$ and $S_t = \varnothing$. Construct a multiset family $\mathcal{T}$ with $|\mathcal{S}| + 2$ copies of $\mathcal{S} \cup \{S_s, S_t\}$, such that $\mathcal{T} = \{S_{ij} : ij \in A\}$ and $S_{ij} = S_i$ for all $ij \in A$ and $i \in V$. Suppose we have subfamily of n multisets $\mathcal{E} \subseteq \mathcal{S}$. Let $V' = \{s,t,i : S_i \in \mathcal{E}\}$. Any simple st-path over V' can be trivially found because the induced subgraph $G[V']$ is complete. We show that any such path is equivalent with respect to the objective. Since $|\mathcal{E}| = n$, $|V'| = n+2$ and the path will have n intermediate vertices. For the path's subgraph $P = (V', A')$, $\{S_{ij} \in \mathcal{T} : ij \in A'\} = \{S_i \in \mathcal{S} : i \in V'\} = \mathcal{E}$. Then, $\uplus_{ij \in A'} S_{ij} = \uplus_{S_i \in \mathcal{C}} S_i$. Thus, if we define $C = \{c_m : m \in \mathcal{U}\}$ where $c_m = 1$ for all $c_m \in C$, we can compute the minimum sum over repeat multiplicities by finding a path with the minimum objective **(38)**, where $M = \uplus_{ij \in A'} S_{ij}$. □

A path on the graph $G = (V, A)$ can be interpreted as a network-flow sent through V from a source vertex $s \in V$ to a sink vertex $t \in V$. Binary flow $x_{ij} \in \mathbb{Z}_2$ is assigned so that $x_{ij} = 1$ if arc ij is in the path. The following three path constraints ensure the balance of inflow and outflow, and distinct arcs respectively:

$$\sum_{j \in V} x_{ji} - \sum_{j \in V} x_{ij} = \begin{cases} -1, & i = s \\ 1, & i = t \\ 0, & i \in V \setminus \{s,t\} \end{cases}, \qquad x_{ij} \in \mathbb{Z}_2 \quad \forall ij \in A. \tag{40}$$

We can easily represent the codon variants of a protein as st-paths in a *multipartite chain*. In this graph, fixed-length sequences of adjacent arcs connect fixed-

length sequences of vertices between s and t in a directed acyclic graph (DAG). We can partition arcs based on their $3(k+1)$- mers' locations in the coding sequence such that arcs are only adjacent if they reside in adjacent partitions. With the multipartite chain, the CODON SCRAMBLING problem clearly becomes intractable for long sequences because graph size scales linearly with the length of the sequence. Notice that the repetitive protein coding sequence is composed of concatenated motifs. We exploit redundancy by constructing instead, a *multipartite cycle* digraph. In this graph, there are instead, r arc-partitions, where r is the number of residues in the repeated motif, a fraction of the protein length. The vertices in the end are connected to the vertices in the beginning by the last arc-partition Figure 13c. Here, the following constraint must be employed to ensure a fixed path length:

$$\sum_{ij \in S} x_{ij} = p - 1, \tag{41}$$

where S is the last arc-partition and p is the number of motif repetitions. This constraint ensures that we return to the beginning of the motif exactly $p - 1$ times.

In order to eliminate disjoint cycles from the set of feasible solutions, subtour elimination constraints must also be used. One such classic strategy, known as the Miller Tucker Zemlin (MTZ) subtour elimination constraints (*53*), offers a concise MILP formulation of the NP-hard Traveling Salesman Problem by introducing an extra variable u_i for each intermediate vertex $i \in V \setminus \{s,t\}$,

$$\begin{aligned} u_i - u_j + q x_{ij} &\le q - 1 \qquad && \forall i \ne j \in V \setminus \{s,t\}, \\ 0 \le u_i &\le q - 1 && \forall i \in V \setminus \{s,t\}, \end{aligned} \tag{42}$$

where q is the scalar number of vertices in a path. Briefly, $u_j \geq u_i + 1$ when $x_{ij} = 1$, forcing an ordering to all vertices that is infeasible in illegal cycles. The linear programming (LP) relaxation of the MTZ constraints is known to be weak (*54*).

By taking advantage of the structure of the De Bruijn graphs, we introduce a modification of the MTZ constraints that give lower bounds on the u variables. In multipartite cycle graphs, every partition is a feedback arc set which contains one arc of every cycle in the graph. Because of this property, the feedback arc set is easily found and the following modified constraints can be used,

$$\begin{aligned} u_i - u_j + px_{ij} &\leq p-1 && \forall ij \in S, \\ u_i - u_j + (p-1)x_{ij} &\leq p-1 && \forall ij \in A \setminus S, i \neq s, j \neq t, \\ 0 \leq u_i &\leq p-1 && \forall i \in V \setminus \ s,t \ , \end{aligned} \tag{43}$$

where $S \subseteq A$ is the designated feedback arc set and p is the scalar number of motif repetitions in the amino acid sequence. Here, MTZ constraints are only enforced at one partition because it is a feedback arc set. For all other arcs, $u_j \geq u_i$ when $x_{ij} = 1$ so that u is maintained along arc ij. This modification allows number of motif repetitions p to replace the number of vertices in a path q in the new constraints. The scalar p always has a significantly smaller value and therefore offers a lower bound on u.

Theorem 2. *Given a digraph* $G = (V, A)$ *and a feedback arc set* $S \subseteq A$*, solutions to the set of constraints in* **(40)** *and* **(43)** *are acyclic st-paths in G.*

Proof. The following proof closely follows the proof from MTZ (*53*). Allow x_{ij} to indicate whether arc ij is in the path of digraph $G = (V, A)$. Suppose we have a cycle of n vertices over vertices $V' \in V$ that contains m arcs from the feedback arc set S. Because S,

when removed from the graph results in a DAG, any path that does not pass S is acyclic and therefore, $m \geq 1$. If we sum the first two inequalities of **(43)** for all $x_{ij} = 1$ on a path over V',

$$\sum_{i \in V'} u_i - \sum_{j \in V'} u_j + (p)m + (p-1)(n-m) \leq (p-1)n \qquad \textbf{(44)}$$

or

$$(p-1)n + m \leq (p-1)n \qquad \textbf{(45)}$$

which is a contradiction since $m \geq 1$. Conversely, consider an acyclic path. When $x_{ij} = 0$, $u_i - u_j \leq p-1$ always since $u_i \leq p-1$ and $u_j \geq 0$. When $x_{ij} = 1$, we can first choose $u_i = k$ where arcs from S have been visited k times to reach vertex i. If $ij \in S$, then $u_i - u_j = -1$ and the first inequality is satisfied. If $ij \in A \setminus S$, $i \neq s$, and $j \neq t$, then $u_i - u_j = 0$ and the second inequality is satisfied. Finally, with the constraints **a)**, the outdegree of s is one, and the indegree of t is one. Because the path is acyclic and due to the constraints in **)**, arcs and vertices are distinct and connected in G. Therefore, solutions are valid st-paths. □

Prior to optimization, it is necessary to choose the De Bruijn window length k to be sufficiently large. The codon-length k places an upper bound x on the nucleotide-length of substrings in ◎. We observe a rule of suffix trees(*55*) that can be applied here: if all substrings of length x do not repeat, all substrings of length $y > x$ do not repeat. In other words, if x is chosen so this condition is satisfied, substrings of length y do not contribute to the objective value and can be properly omitted from $\mathcal{U}$.

In the k-mer De Bruijn subgraph, each vertex traversal guarantees that a substring $m \in \mathcal{U}$ of length $3k$ is encountered. Notice that constraints **(43)** ensure that vertices in paths are distinct. It is then reasonable to set $x = 3k$, where all substrings $m \in \mathcal{U}$ have length $\leq x$. The condition that x does not repeat can be most simply ensured with the set packing constraint,

$$\sum_{ij \in A} w_{ij}^{m} x_{ij} \leq 1, \tag{46}$$

for all $m \in L$ where $L \subseteq \mathcal{U}$ is the subset of substrings of length $3k$, and the parameter w_S^m is the nucleotide multiplicity of substring m in arc ij.

It is also clear that there is an advantage to choosing a small codon-length k: it has an exponential relationship with the cardinality of A. That is, the total number of arcs in the multipartite cycle graph is

$$\sum_{i=1}^{r} \prod_{j=1}^{n+1} a_{ij} \tag{47}$$

where a_{ij} is the number of redundant codons for the residue at position j for arc-partition i, and r is the number of residues in the motif. We calculate a lower bound for k in the annotated supplementary code (https://chilkotilab.pratt.duke.edu/codon-scrambler). First, we define *codon-diversity* as the number of possible distinct synonymous sequences for a given length, peptide subsequence and reading frame. Briefly, we find the minimum k for a coding sequence that satisfies the following condition: for each $3k$-length reading frame of every distinct peptide subsequence, its

codon-diversity must exceed its number of occurrences in the full peptide sequence. Otherwise, the codon-diversity is exhausted and 3*k*-mers must definitely be repeated.

Finally, we combine the constraints **(39)**, **(40)**, **(41)**, **(43)**, and **(46)** in the final MILP formulation (Table 1). Besides tandem-repetitive proteins, we also realize the case when motifs are repeated but interspersed. Here, given that peptide motifs are separated by non-repetitive segments, we can optimize the set of unconnected motifs. Briefly, we create the previously defined multipartite chain with *r* arc partitions, where *r* is the number of residues in the motif. Its MILP formulation is consistent with the standard case, except that constraint **(43)** is unnecessary for a DAG, and constraint **(40)** is modified so that instead of one unit of flow, multiple units are sent from source *s* to sink *t*.

3.4 Computational analysis for optimization of repetitive proteins

To test the algorithm, we selected a diverse set of 19 repetitive proteins. The majority of these proteins are bioinspired from fibrous elastomeric animal proteins (*56*–*61*) ranging from highly resilient proteins —elastin, abductin, and resilin— to the highly tough silk fibroin. We chose other repetitive proteins for their broad array of applications: peptide drugs (*62*), ligand scaffolds (*63*), purification tags (*20*, *64*), protein binders (*65*, *66*), and hydrogel-forming proteins that adopt β-sheet (*67*, *68*), coiled-coil (*68*), or random coil (*69*) structures. These repetitive protein sequences can be completely optimized with CPLEX; computational experiments (Table 2) show that global optimality is reached within times ranging from 460 ms for wheat gliadin (SPR16) (*70*) to 2,020 s for elastin like polypeptide (ELP) (*71*).

Table 2: Computational results for the optimization of repetitive proteins

	Sequence	Objective value	Time (s)
Designed ankyrin repeat protein (Darpin E_01) (*65*)	DLGKKLLEAARAGQDDEVRILMANGADVNA DDTWGWTPLHLAAYQGHLEIVEVLLKNGADVNA YDYIGWTPLHLAADGHLEIVEVLLKNGADVNA SDYIGDTPLHLAAHNGHLEIVEVLLKHGADVNA QDKFGKTAFDISIDNGNEDLAEILQ	1.65×10^3	684
Glucagon-like peptide (GLP-1) (*62*)	$[GAHGEGTFTSDVSSYLEEQAAKEFIAWLVKGR]_6$	7.09×10^3	10.9
Adenovirus-like construct (Ad I) (*72*)	$[LSVQTSAPLTVSDGK]_{14}$	1.20×10^4	72.2
Mussel adhesive protein (MAP) (*57*)	$[AKPSYPPTYK]_{16}$	3.41×10^4	1.33
Repeats in toxin (BRT17) (*64*)	$[GGAGNDTLY]_{17}$	5.04×10^4	31.3
Wheat gliadin (SPR16) (*70*)	$[PQQPY]_{16}$	9.23×10^4	0.460
Cell adhesive substrate (poly-RGD) (*66*)	$[GSGSGSGRGDS]_{20}$	1.05×10^5	79.5
β-sheet forming polypeptide (poly(alanylglycine)) (*68*)	$[AGAGAGPEG]_{10}$	1.22×10^5	5.06
Resilin like polypeptide (Dros16) (*58*)	$[GAPGGGNGGRPSDTY]_{16}$	1.74×10^5	603
Abductin (AB12) (*59*)	$[FGGMGGGNAGFGGMGGGKAGFGGMGGGNAG]_4$	2.00×10^5	24.6
Transglutaminase substrate peptide (BQ_6) (*69*)	$[[GQQQLGGAGTGSA]_2[GAGQGEA]_3]_6$	3.44×10^5	285
Silk-elastin like polypeptide (SELP0K) (*60*)	$[[GAGAGS]_2[GVGVP]_4GKGVP[GVGVP]_3]_6[GAGAGS]_2$	4.72×10^5	913
Alanine-rich polypeptides (35-H-6) (*63*)	$[AAAQAAQAQAAAEAAAQAAQAQ]_6$	8.11×10^5	67.35
Elastin like polypeptide ($ELP[V_5A_2G_3$-60]) (*71*)	$[[GVGVP]_2GGGVPGAGVP[GVGVP]_3GGGVPGAGVPGGGVP]_6$	9.77×10^5	2,020

β-sheet forming polypeptide (poly-EAK9)	[AEAEAKAK]$_{18}$	1.16×10^6	2.45
Collagen like protein (CLP3.7) (*61*)	[GAPGTPGPQGLPGSP]$_{24}$	1.22×10^6	18.0
Elastin like polypeptide (ELP[AV-60]) (*20*)	[GAGVPGVGVP]$_{30}$	1.35×10^6	14.1
Elastin like polypeptide (ELP[V-60]) (*71*)	[GVGVP]$_{60}$	1.48×10^6	4.02

The optimization objective eliminates repeats, while favoring shorter repeats with weak hybridization energies. In contrast, if two or more oligonucleotides are concatemerized (*56*), the resulting coding sequence is highly repetitive. This is apparent from the identity dot plot of ELP, where long intense lines off the main diagonal indicate long repeats (Figure 14). After optimization, these lines disappear or break into smaller lines with nucleotide lengths of ≤8 bp.

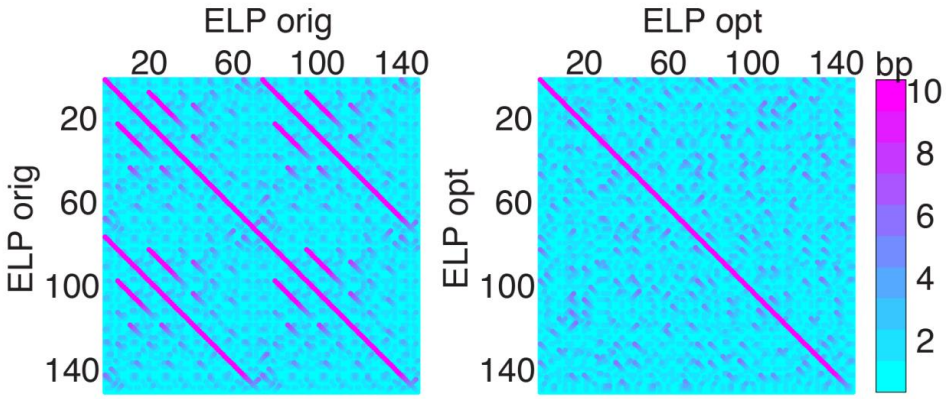

Figure 14: Nucleotide sequence dot plot. 10 repeats of the GVGVP motif in ELP compared against itself with codons that would result from concatemerization with two oligonucleotide inserts (ELP orig, left) and with optimal non-repetitive codons (ELP opt, right). Points describe exact sequence identity, including the reverse complement. From (*40*).

Computational experiments show how CPLEX timings scale with the number of repeat units of the ELP motif (Figure 15). As expected, computational run times have an exponential relationship with the increasing maximum repeated nucleotide length. This nucleotide length determines the De Bruijn window length *k*, which is exponentially proportional to the overall size of the graph and MILP problem. In contrast, the objective function increases continuously with the increasing number of motif repeats. For reasonable values of *k*, experimentally-useful numbers of repeats can be readily optimized (Table 2).

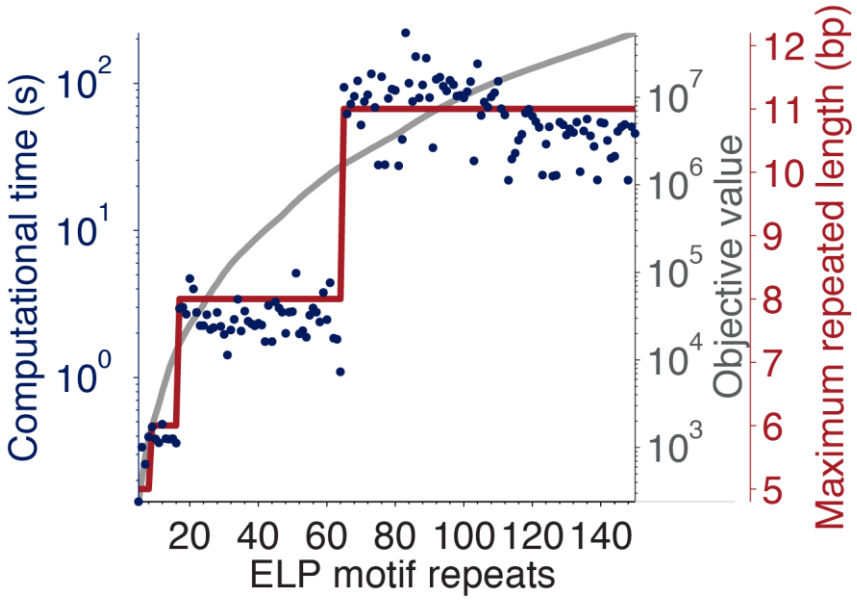

Figure 15: Timings of CPLEX optimization. Optimization was carried out for ELP with up to 150 motif repeats, with corresponding objective values and maximum repeated nucleotide lengths. From (*40*).

3.5 Codon scrambling enables gene assembly with PCR

To illustrate the broad applicability of codon scrambling, we show that optimization enables gene synthesis of a diverse set of repetitive coding sequences. First, we selected 5 of the optimized genes because they span a range of objective values: GLP-1, BRT17, ELP[V-30], AB12, and 35-H-6. Using polymerase cycling assembly (PCA), we assembled their optimized sequences and their random-codon variants from their corresponding 80–90 nt oligonucleotide building blocks. For GLP-1, we additionally assembled a purely repetitive nucleotide sequence that would otherwise result from concatemerization of a gene that encodes a monomer of GLP-1. Microfluidic electropherograms of the resulting PCA products show many nonspecific bands for not only the purely repetitive codon variant, but also the random-codon variants (Figure 16a). Due to the lack of distinct bands, we were unable to sequence or clone these genes. On the other hand, the optimized codon variants had distinct bands, and could be cloned into the pET-24a(+) expression vector and sequence verified (Table 3).

Table 3: Gene synthesis error rates

Gene	Fragment length (bp)	Clones sequenced	Error-free clones	Total deletions*	Total insertions*	Total substitutions*
GLP-1	644	5	2	5	1	0
BRT17	731	5	1	5	0	0
ELP[V-30]	510	5	1	6	0	0
AB12	425	5	2	4	0	0
35-H-6	464	5	2	8	0	1

*Defects were summed over all sequencing chromatograms containing errors.

Gen9 Inc. and Genscript Inc. successfully synthesized the other 14 optimized sequences (Figure 16b). In contrast, non-optimized random-codon variants were far too repetitive to pass their sequence complexity filters.

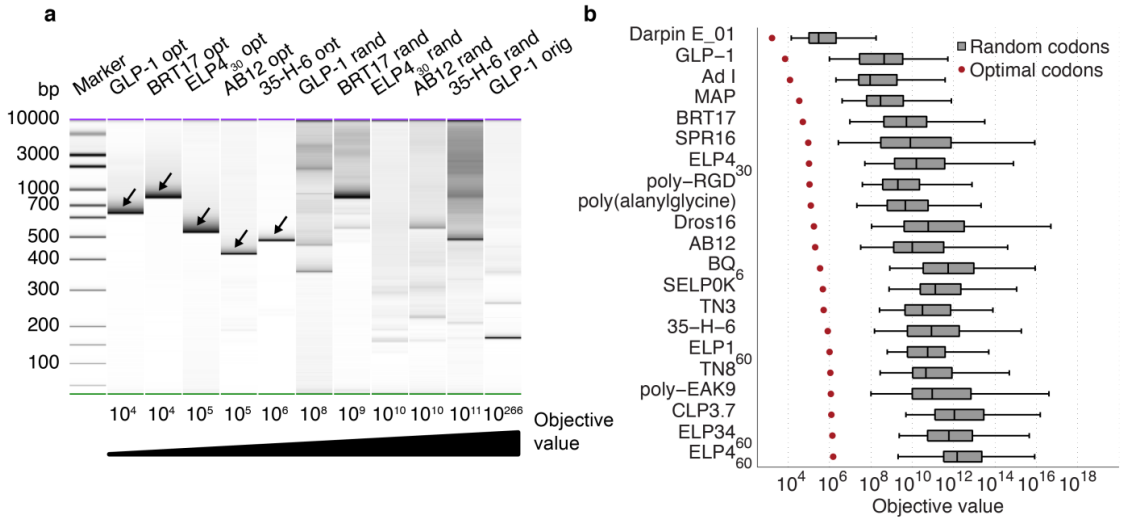

Figure 16: Gene assembly of a diverse set of repetitive proteins. a, Microfluidic electropherograms of PCA products, rank-ordered by calculated objective values. We assembled oligonucleotides into full-length genes with optimal codons (opt), randomized codons (rand) and original codons that would otherwise result from concatemerization (orig). The distinct bands marked by arrows represent assembled genes of correct length and sequence that are subsequently cloned into the pET-24a(+) vector. **b**, Rank-ordered objective values of the optimal-codon variant and 100 random-codon variants of each repetitive gene. Boxes signify the interquartile ranges (IQR) between the 25th and 75th percentiles and the lines signify the medians. From (*40*).

Additionally, we show that the codon scrambling approach can also be applied to repetitive proteins with variable regions like transcription activator-like effectors (TALEs) and native semi-repetitive proteins like spider-silk, as their genes are also commonly rejected from gene synthesis vendors and require even more specialized methods of synthesis (*73*, *74*). The optimized sequences for these genes, TN3 (*75*) and ADF-1 (*74*), were successfully synthesized and sequenced.

3.6 Codon scrambling enables Golden Gate and Gibson Assembly

A rank-order of objective values from Figure 2a suggests that a soft threshold between 10^6 and 10^7 exists between success and failure of PCR-based techniques. Longer sequences with objective values greater than 10^7 run the risk of PCA failure. This limitation can be addressed by a wide array of PCR-based manipulations of codon-scrambled genes that facilitate the assembly of longer multi-kilobase genes from shorter ones in a single cloning step. To illustrate this, we show that Type IIS restriction sites can be introduced into the 5′ flanking sequence of PCR primers so that they are integrated into PCR products, which are then assembled together with the Golden Gate assembly method to create even longer repetitive genes. A Golden Gate reaction can assemble as many as 10 DNA fragments in a linear order where approximately 90% of recombinant clones contain the desired construct(*76*). We used this strategy to create 150 repeats of the GVGVP amino acid motif (ELP[V-150]), which had a minimum possible objective value of 5.4×10^7, from optimized and PCA-assembled monomers consisting of 30 repeats (ELP[V-30]) (Figure 17a, b). After PCR and Golden Gate assembly, DNA restriction analysis confirmed that an estimated 93% of recombinant clones contained the desired construct (Figure 17c).

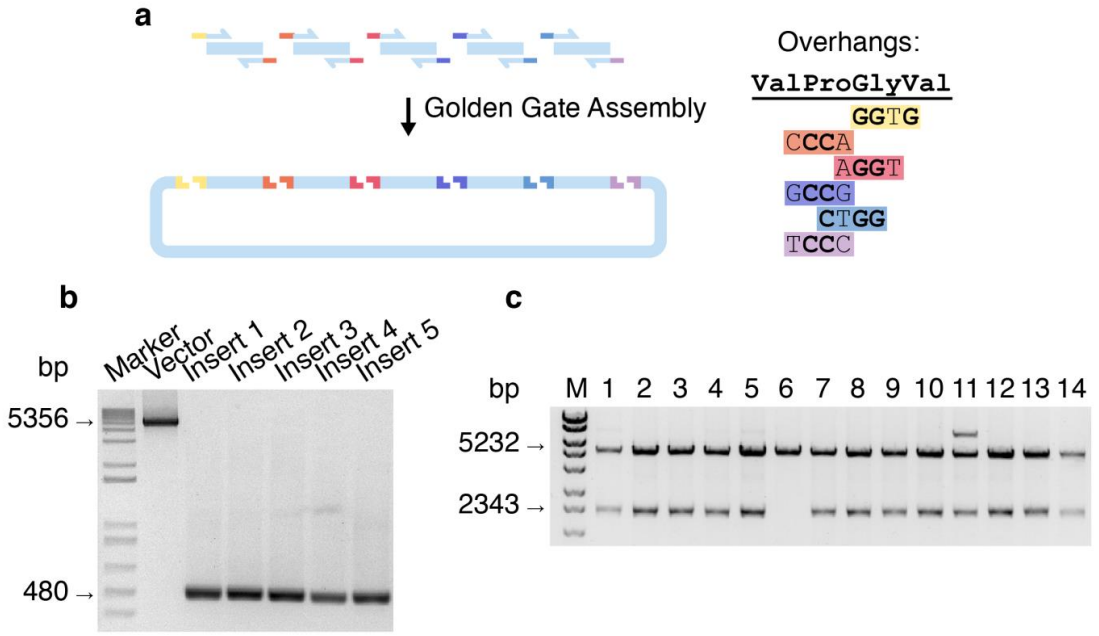

Figure 17: Golden Gate assembly of ELP[V-150]. a, Schematic of Golden Gate assembly of the gene coding ELP[V-150] from PCR-amplified ELP[V-30] inserts. Primers attach unique scar-less 4 bp junctions and flanking BsaI sites to ELP[V-30] fragments. The Golden Gate digestion-ligation reaction joins these PCR products in a linear order based on the unique junctions. **b**, PCR products from ELP[V-150] construction. The linearized pET-24a(+) vector should have a length of 5356 bp. The amplified ELP[V-30] inserts should have lengths of ~480 bp. **c**, Diagnostic digest of Golden Gate assembled ELP[V-150]. When ELP[V-150] is fully assembled and digested from the vector with XbaI and BamHI, it should have a length of 2343 bp. Agarose gel electrophoresis shows that of the 14 plasmids extracted from individual colonies, one does not contain an insert (lane 6). From (*40*).

The benefits of this method over recursive ligation methods(*77*) become greater with longer target gene lengths (Table 4). For example, the synthesis of genes with lengths of 5 kbp require only two cloning steps by codon scrambling, while recursive ligation requires six cloning steps.

Table 4: Rounds of cloning to assemble full-length genes

	Cloning steps	
Gene length (bp)	Recursive ligation	Codon scrambling
150	1	1
300	2	1
600	3	1
1200	4	1
2400	5	2
4800	6	2

We further demonstrate that the optimized genes, obtained from PCA or commercial synthesis, could be PCR-amplified with gene-specific primers for Gibson cloning into new vectors (*78*). Microfluidic electropherograms of the PCR products (Figure 18) show few to no nonspecific bands. We seamlessly cloned these PCR-amplified genes from their original pET-24a(+) vectors into linearized pMAL-c5X vectors containing a C-terminal His_6-tag and an N-terminal Maltose Binding Protein (MBP) tag, for enhanced soluble expression in Escherichia coli. Gene-specific Gibson assembly primers were designed to amplify the coding sequence (CDS) of each gene and add overhangs containing the vector-overlapping sequences, 5′-GAAGGTCTTCCGGT-3′ and 5′-GGCTCGGGACATCATC-3′. All plasmids are available from Addgene (Plasmids #66996–#67015).

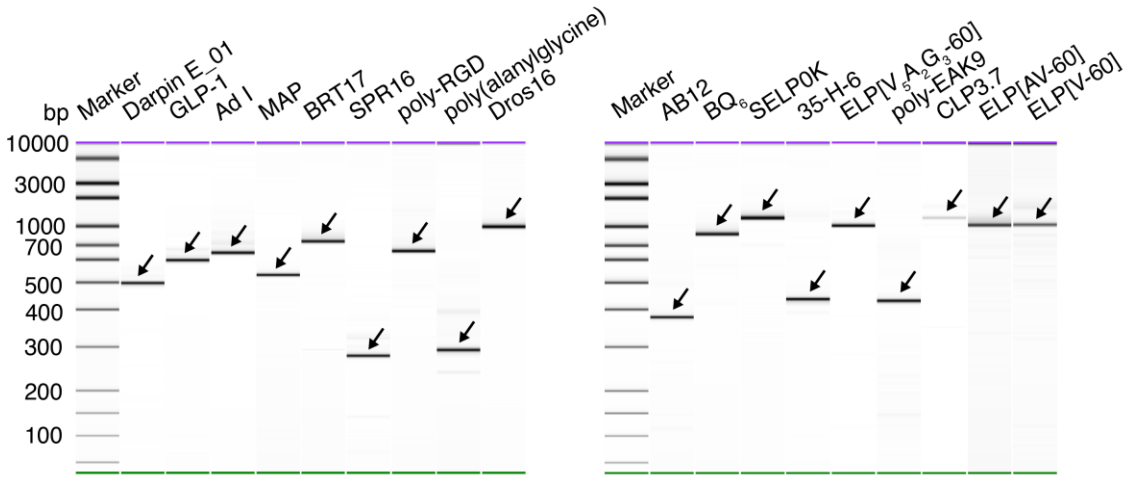

Figure 18: Microfluidic electropherograms of PCR products of optimized genes. Each PCR reaction successfully amplified the gene with gene-specific primers for subsequent Gibson cloning. The distinct bands marked by arrows represent PCR products of correct lengths and sequences. From (*40*).

3.7 Effects of codon scrambling on recombinant protein expression

We finally show that manipulating codon usage with our method does not stifle heterologous expression of these repetitive coding sequences in E. coli, as there is a concern that synonymous codon usage is known to affect protein expression levels (*79*). Despite the omission of codon bias rules during the codon scrambling of repetitive proteins, we found sufficient expression of 18 MBP-tagged genes for downstream characterization experiments (Figure 19a). We further confirmed full-length products by the presence of C-terminal His_6-tags via western blot analysis (Figure 19b) and direct DNA sequencing from E. coli cultures.

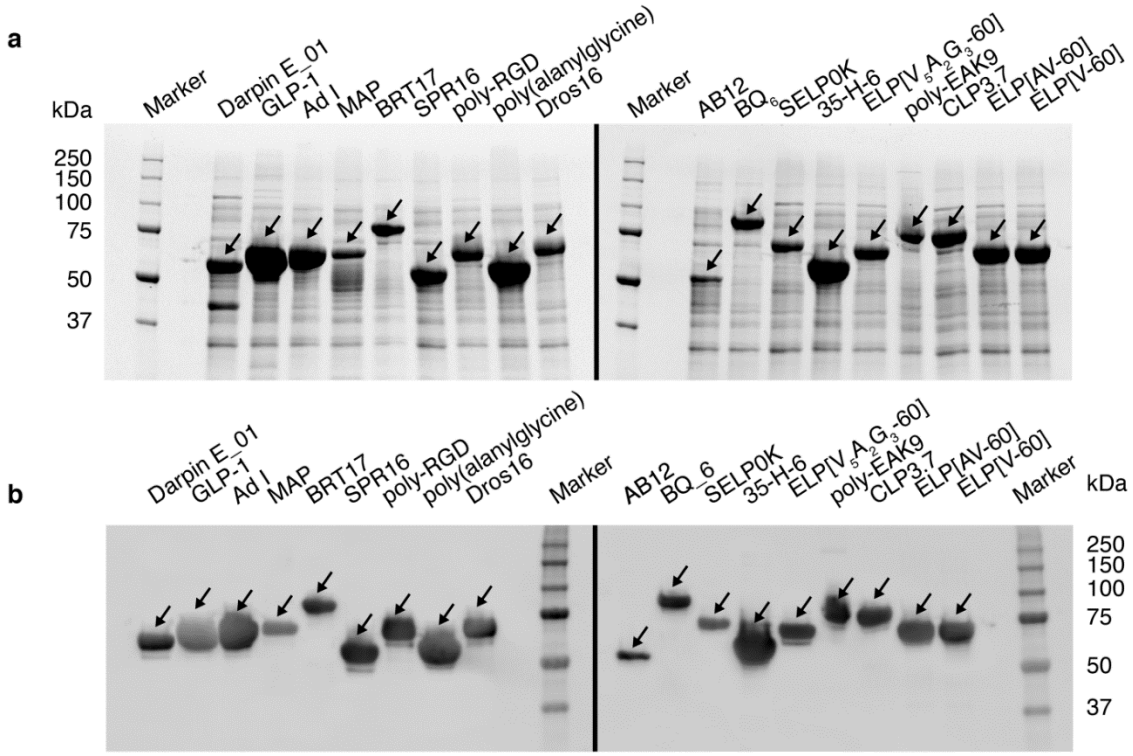

Figure 19: Protein expression of a diverse set of repetitive proteins. a, Recombinant expression of MBP-tagged repetitive proteins in E. coli from optimized gene sequences. 50 μg of total cell lysate for each protein were separated and visualized on tris-glycine extended (TGX) stain-free gels. The broad bands marked by arrows represent expressed repetitive proteins. **b,** Western blot analysis of expressed repetitive proteins. 50 μg of total cell lysate for each protein were first separated on TGX stain-free gels. The distinct bands marked by arrows represent full-length repetitive proteins with C-terminal His6-tags. From (*40*).

We also directly compared the protein yields of ELP[V-30] with codons optimized for assembly (ELP[V-30] opt) with the same ELP with highly repetitive codons that were selected based on *E. coli* codon usage bias (ELP[V-30] orig) (Supplementary Fig. 4)(*56*, *79*). Relative quantification by gel densitometry suggests that the expression of ELP[V-30] opt is at least 80% that of ELP[V-30] orig.

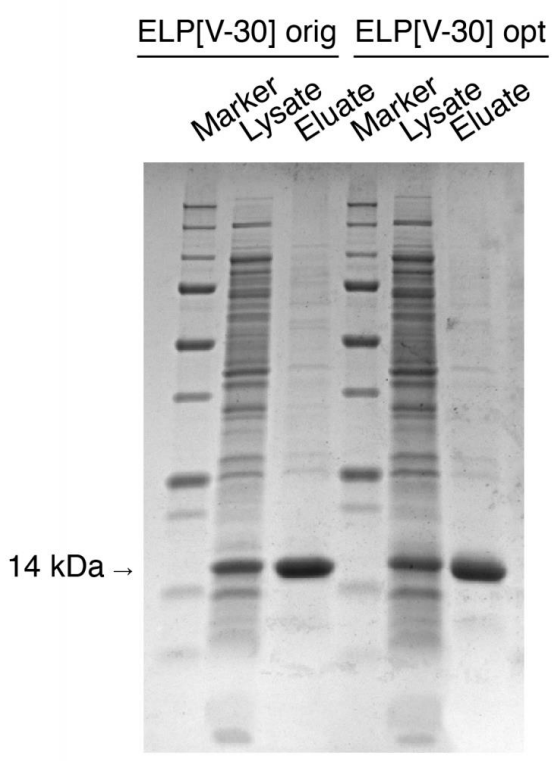

Figure 20: Effects of codon scrambling on protein expression and purification. We expressed and purified His6-tagged ELP[V-30] with codons optimized for assembly (ELP[V30] opt) and the same ELP with codons that were selected based on E. coli codon usage bias (ELP[V-30] orig). Following purification, 30 μg of total cell lysates and eluates were separated by SDS-PAGE on a TGX precast gel and then stained with EZBlue gel staining reagent (Sigma). Band intensities were quantified in Image Lab software (BioRad) and normalized by total protein. Relative quantification by gel densitometry shows that ELP[V-30] opt achieves protein yields of approximately 80% of ELP[V-30] orig. From (*40*).

Translation in bacteria involves initiation, elongation, termination and ribosome turnover. At the same time, interactions with the expression environment vary protein solubility, mRNA stability, protein localization, and cell viability. A number of theories have been proposed in the past to codon optimize genes. Some of the traditional schemes offered at various gene synthesis vendors include CAI (codon adaptation index) maximization, codon sampling, dicodon optimization, and codon frequency matching (*80*). Other theories include the importance of number of rare codons, pairs of

consecutive rare codons, frequency of rare-pairs, number of potential RNaseE cleavage sites, Shine-Dalgarno occlusion by secondary structure, number of predicted rho-independent transcription termination signals, and many others (*81*). Kudla *et al.* shed light on the question of codon optimization for the purposes of synthetic gene expression. They performed a systematic study, where they generated a library of 154 GFP codon variants in which the codon usage was varied randomly at synonymous sites. They expressed these proteins from a standard expression vector in *E.coli*, for a 250-fold variation in protein levels, as measured by GFP intensity in a microplate reader. The stability of mRNA folding near the start codon explained more than half the variation in protein levels. Secondary structures can inhibit rates of translation initiation, and was the dominant factor in influencing gene expression levels. Codon variation can vary mRNA free energies. The authors concluded that the key factor is avoidance of secondary structures near the ribosomal binding site, and that codon bias is important in fewer cases than previously thought.

Our findings may be due to presumably balanced codon usage which do not include many offending codons overall, as well as the presence of a cleavable N-terminal tag with low mRNA folding energy, which promotes efficient translation initiation (*82*). Notably, however, the initial expression of AB12 only produced truncated proteins. We located a high percentage of ATG and Shine-Dalgarno (SD)-like motifs in the AB12 gene, which suggested that the poor expression of AB12 was caused by recurrent iRBS (Figure 21a, b). Recoding the gene to eliminate the recurrent iRBS restored expression levels likely due to the avoidance of translational stalling (*50*).

To demonstrate the utility of codon scrambling for mammalian cell expression, HEK293T human cells were transfected with codon-scrambled transcription activator-like effector nuclease (TALEN)-encoding plasmids produced in this study, which were designed to target exon 51 in the human dystrophin gene and correct its disrupted reading frame in Duchenne muscular dystrophy (DMD) patients (*75*). TALENs are nonspecific nucleases fused to an engineered DNA-binding domain. Following transfection, TALENs mediate genetic correction by non-homologous end joining (NHEJ). This is detected by the Surveyor assay, which identifies insertions or deletions by denaturing and reannealing mutant alleles and using the Cel I nuclease to cleave the resulting mismatches. The resulting cleavage products are apparent by polyacrylamide gel electrophoresis (PAGE). Both TALENS containing codons optimized for assembly and original codons (*75*) modified exon 51, as detected by the Survey assay (Figure 21c).

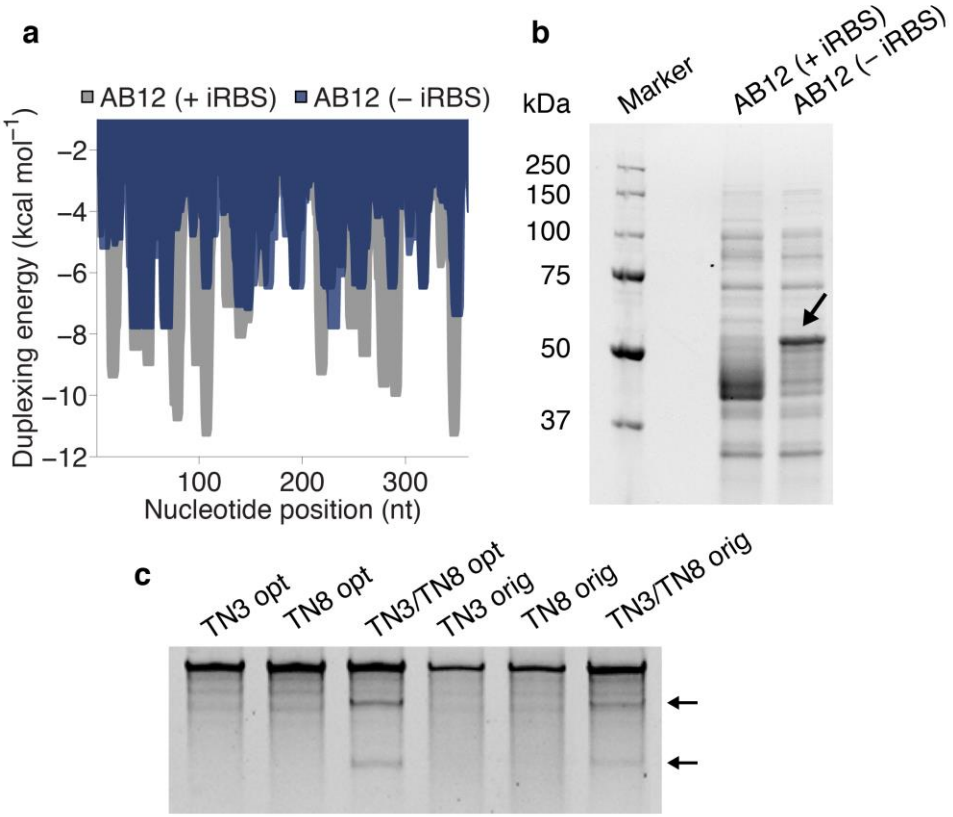

Figure 21: Expression and utilitization of repetitive proteins. a, Sliding window analysis to identify iRBS motifs. Minimum duplexing energies of 15-nt windows with the *E. coli* anti-SD sequence are plotted against their nucleotide position from the start of the AB12 CDS. We recoded AB12 (– iRBS) to eliminate windows with values lower than –8 kcal mol^{-1}. **b**, Comparison of MBP-tagged AB12 with and without iRBS. 50 μg of total cell lysate for each protein were separated and visualized on TGX stain-free gels. Expression of AB12 (+ iRBS) produced a smear of truncated products. Expression of AB12 (– iRBS) produced a broad band of full-length protein, marked by an arrow, with the expected molecular weight of 53.4 kDa. **c**, CelI detection of human genome editing mediated by the optimized TALEN pair TN3/TN8. HEK293T cells were transfected with TALEN plasmids containing optimal non-repetitive codons (opt) and original codons that result from the Golden Gate assembly method (orig) (*73*). NHEJ activity was detected with the Surveyor assay; the bands on the tris-borate-EDTA (TBE) polyacrylamide gel marked by arrows represent the expected cleavage fragments of the PCR-amplified target locus (*75*). From (*40*).

Finally, as a test of protein functionality, different codon-scrambled ELP genes in the pET-24a vector were expressed and His_6-tag-purified, and exhibited the expected

thermally responsive phase separation that is characteristic of this class of protein polymers (*20*).

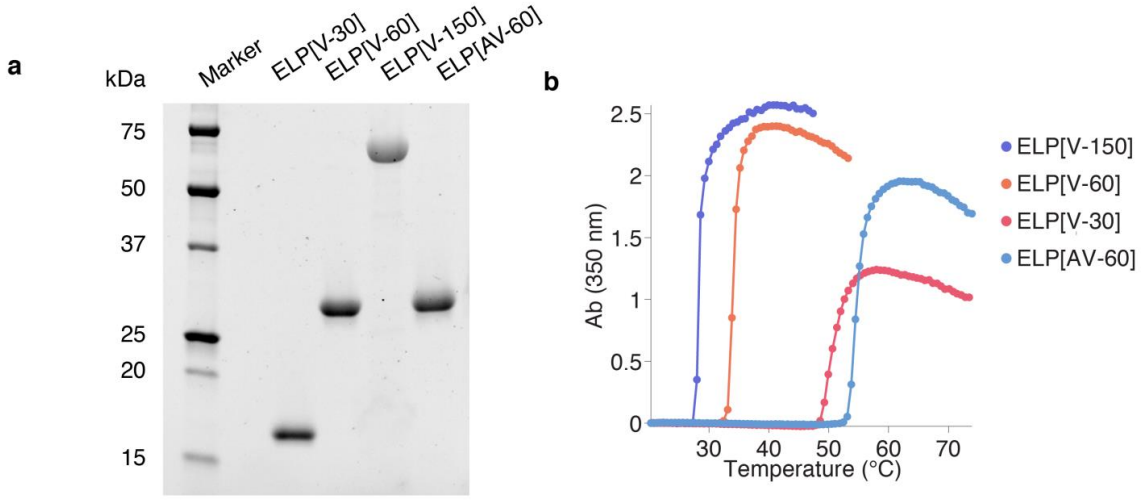

Figure 22: Purification of ELPs. a, 150 pmol of each purified ELP were separated on TGX stain-free gels. **b**, Turbidity profiles for purified ELPs with inverse phase transition behavior. Observed transition temperatures occur at 28.6 °C, 35.2 °C, 50.6 °C, and 55.2 °C for ELP[V-150], ELP[V-60], ELP[V-30], and ELP[AV-60] respectively. These transition temperatures are within 1 °C of model-predicted temperatures: 28.6 °C, 34.8 °C, 50.4 °C, and 55.8 °C (*20*). From (*40*).

In conclusion, this strategy provides a robust and general solution to the long-standing problem of the scalable synthesis of highly repetitive coding sequences that are amenable to further oligomerization and manipulation. The open access availability of the gene design and optimization software that can be implemented on a personal computer will enable the rapid design of genes for a wide range of repetitive proteins that are easily amenable to further manipulation owing to their optimized, minimally repetitive sequence — using powerful gene manipulation techniques such as Gibson assembly and PCR amplification. This design methodology also provides the means to systematically design and synthesize new stimuli-responsive proteins of interest.

3.8 Materials and methods

3.8.1 Gene design and assembly with PCA

Genes were designed so that the coding region was flanked by two BsaI recognition sites. BsaI generates 5′ overhangs for the purpose of directional scarless cutting in Golden Gate cloning. This entire sequence was additionally flanked by forward and reverse universal primer recognition sequences for the purpose of downstream PCR amplification. Overlapping oligonucleotides were designed with GeneDesign(*83*) with a target overlap-length of 40 bp, assembly length of 90 bp, and T_m of 70 °C. 1.25 pmol each of 8–12 oligonucleotides (Integrated DNA Technologies) of approximately 90 bp length were assembled into full-length genes in a 50 μL reaction containing 250 μM of each deoxynucleotide triphosphate (dNTP), 1X Herculase II PCR buffer, and 0.5 μL Herculase II Fusion polymerase (Agilent Technologies). The reaction was incubated at 98 °C for 15 min, followed by 15 cycles at 98 °C for 20 s, 57–67 °C for 20 s and 72 °C for 30 s, and a final step at 72 °C for 3 min. 1 μL of the full-length PCA product was then amplified by PCR using terminal universal primers. The 50 μL PCR reaction contained 250 μM of each dNTP, 1X Herculase II PCR buffer, 0.25 μM each of universal forward and reverse primers and 0.5 μL Herculase II Fusion polymerase. The reaction was incubated at 98 °C for 15 min, followed by 15 cycles at 98 °C for 20 s, 64 °C for 20 s and 72 °C for 30 s, and a final step at 72 °C for 3 min. The product was diluted 1:50 and then visualized with microfluidic electrophoresis using the Agilent high sensitivity DNA kit and the 2100 Bioanalyzer instrument (Agilent Technologies). BsaI recognition sites, 4 bp overhangs (CGGT/ GGCT), and the His_6-tag coding region were

added to a modified pET-24a(+) vector using PCR. The construction of this modified vector from pET-24a(+) (Novagen) has been previously documented(*56*). 2.5 ng of the plasmid was linearized and amplified in a 50 μL PCR reaction with the same components as previously described. The reaction was incubated at 95 °C for 2 min, followed by 5 cycles at 95 °C for 20 s, 56 °C for 20 s and 72 °C for 2 min 40 s, followed by 25 cycles at 95 °C for 20 s and 72 °C for 3 min, and a final step at 72 °C for 3 min. The gene assembly product and linearized vector were gel-purified from a 1% and 0.4% low melting point agarose gel respectively using the QIAquick gel extraction kit (Qiagen). The concentration of the linearized vector was adjusted to 100 ng μL^{-1}, and gene assembly products were adjusted to a 1:1 molar ratio to the vector. 1 μL of each purified product was combined in a 10 μL Golden Gate digestion-ligation reaction containing 0.75 μL BsaI (20 U μL^{-1}; New England Biolabs), 1X NEBuffer 4, 1X bovine serum albumin (BSA), 0.25 μL T7 Ligase (3000 U μL^{-1}; Enzymatics), and 1 mM adenosine triphosphate (ATP). The reaction was incubated for 20 cycles at 37 °C for 5 min and 20 °C for 5 min. The 7 μL of the Golden Gate product was treated with exonuclease to remove incomplete ligation products in 10 μL reaction consisting of 1 μL PlasmidSafe DNAse (10 U μL^{-1}; Epicentre), 1X PlasmidSafe Reaction Buffer, and 1 mM ATP. The reaction was incubated at 37 °C for 30 min, followed by inactivation at 70 °C for 30 min. 5 μL of the PlasmidSafe reaction product was transformed into 50 μL EB5α competent cells (EdgeBio) following the manufacturer's instructions. Successful clones were screened with direct sequencing from colonies using standard T7 primers. Plasmids

were prepared from sequence-verified colonies using the QIAprep spin miniprep kit (Qiagen).

3.8.2 PCR and Golden Gate assembly of ELP

BsaI recognition sites, 4 bp overhangs (GGAG/ACCC), and the His_6-tag coding region were added to the modified pET-24a(+) vector using PCR. ELP[V-30] was then gene-assembled with PCA and cloned into the modified pET-24a(+) vector as described in the previous section. 6 pairs of custom primers were designed to attach BsaI recognition sites with unique 4 bp overhangs so that all 5 PCR products and a linearized modified pET-24a(+) vector would ligate together in a single Golden Gate reaction to form the gene coding ELP[V-150]. Each 4 bp overhang was designed to be unique using codon degeneracy and/or 2 bp shifts in either direction (Figure 17a). 2.5 ng of ELP[V-150]-pET24a plasmid was amplified in 50 μL PCR reactions with the same components as previously described. The reactions were incubated at 98 °C for 3 min, followed by 30 cycles of 98 °C for 20 s, 61 °C for 20 s and 72 °C for 10 s, and a final step at 72 °C for 3 min. The PCR products were visualized and gel-purified from a 2% low melting point agarose gel using the QIAquick gel extraction kit. The concentration of each insert was estimated from the gel using the Image Lab software (BioRad). The inserts were each adjusted to 2.68 ng μL^{-1}, or 1:1 molar ratio with 100 ng μL^{-1} linearized vector. 1 μL of each of the 6 gel-purified PCR products were then assembled in a 10 μL Golden Gate digestion-ligation reaction, treated with PlasmidSafe exonuclease, and transformed into EB5α competent cells with components and incubation times as described previously.

Plasmids were prepared from 14 colonies, and were subjected to DNA restriction analysis with XbaI and BamHI, followed by DNA sequencing.

3.8.3 ELP protein expression and purification

Modified pET-24a(+) plasmids containing ELP[V-30], ELP[V-60], ELP[V-150] and ELP[AV-60] were each transformed into BL21 Star competent cells (Life Technologies) following the manufacturer's instructions. Colonies were inoculated in 2 mL of Terrific Broth (TB) plus 100 μg mL^{-1} kanamycin in 14 mL round-bottom tubes and grown for 8 hours at 37 °C and 200 rpm. 100 μL of the starter cultures were inoculated in 10 mL of Overnight Express TB (EMD Millipore) plus 100 μg mL^{-1} kanamycin in 50 mL Bio-Reaction tubes (Celltreat) and grown for 24 h at 30 °C and 140 rpm. The cells were pelleted and lysed with 0.5 mL B-PER complete bacterial protein extraction reagent (Life Technologies) according to the manufacturer's instructions. 1X Halt protease inhibiter cocktail (Life Technologies) and 1 mM EDTA was added and the lysates were subject to 4 cycles of freeze-thaw. Lysates were purified using the HisExpress columns (ClaremontBio) following the manufacturer's instructions. 10 mL of 1X phosphate buffered saline (PBS) plus 10 mM imidazole was used as the wash buffer and 1 mL of 1X PBS plus 250 mM imidazole was used as the elution buffer. Protein yields were quantified with the Quant-iT protein assay kit (Life Technologies) and were diluted with 1X PBS to 20 μM concentrations. 150 pmol of each ELP was also separated by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) on a Mini-PROTEAN TGX stain-free precast gel (BioRad). The gels were exposed to 1 minute of 302 nm UV light to produce fluorescence and imaged with a Gel Doc XR system (BioRad). To

characterize the inverse transition temperature of ELPs, the 350 nm optical densities of 25 μM ELP solutions were monitored with a Cary 300 UV-Vis spectrophotometer (Agilent Technologies) as a function of solution temperature.

3.8.4 PCR and Gibson assembly

Modified Gene-specific Gibson assembly primers were designed to amplify the coding sequence (CDS) of each gene and add overhangs containing the vector-overlapping sequences, 5′-GAAGGTCTTCCGGT-3′ and 5′-GGCTCGGGACATCATC-3′. 0.25 ng of pET24a(+) plasmids containing each gene were amplified with their respective primers in 50 μL PCR reactions with the same components as previously described. The reactions were incubated at 98 °C for 3 min, followed by 30 cycles of 98 °C for 20 s, 47.7–64.4 °C for 20 s and 72 °C for 30–60 s, and a final step at 72 °C for 3 min. The products were diluted 1:50 and then visualized with microfluidic electrophoresis using the Agilent high sensitivity DNA kit and the 2100 Bioanalyzer instrument. The pMAL-c5X vector (New England Biolabs) was linearized to contain the overlap sequences at each end with primers using PCR. All PCR products were then purified with the AxyPrep Mag PCR clean-up kit (Axygen) and quantified on a NanoDrop 1000 spectrophotometer (Thermo Scientific) by the absorbance at 260 nm. 30 ng of linearized pMAL-c5X and 60 fmol of the insert, at 3.5:1 molar ratio with the vector, were combined in a 10 μL Gibson assembly reaction containing 1X Gibson assembly master mix (New England Biolabs). The reaction was incubated at 40 °C for 1 h. 2 μL of a 1:4 dilution of each Gibson product were transformed into 50 μL EB5α competent cells following the manufacturer's instructions. Successful clones were screened with direct sequencing from colonies using

pMAL sequencing primers. Plasmids were prepared from sequence-verified colonies using the QIAprep spin miniprep kit. All plasmids are available from Addgene (Plasmids #66996–#67015).

3.8.5 Recombinant protein expression

pMAL-c5X plasmids containing each gene were transformed into NEB Express competent cells (New England Biolabs) following the manufacturer's instructions. Colonies were inoculated and grown in 2 mL of TB plus 50 μg mL^{-1} carbenicillin, then inoculated and grown in 10 mL of Overnight Express TB plus 50 μg mL^{-1} carbenicillin, and finally pelleted and lysed as described previously for ELP protein expression. Lysates were diluted based on dry pellet weight to 6.67 μg μL^{-1} with 1X SDS plus 1X Halt protease inhibiter cocktail and 1 mM EDTA. 50 μg of each lysate was heated to 98 °C for 10 min and then separated by SDS-PAGE on Mini-PROTEAN TGX stain-free precast gels. The gels were exposed to 1 minute of 302 nm UV light to produce fluorescence and imaged with a Gel Doc XR system. The presence of the C-terminal His_6-tag was verified with Western blot analysis, using a 1:4000 dilution of DyLight 650 conjugated 6x-His epitope tag antibody (Life Technologies, cat. no. MA1-21315-D650). The resulting western blot was imaged using a Typhoon 9410 (GE Healthcare).

3.8.6 Computation

The minimum hybridization energies for all substrings of each repetitive protein sequence were calculated with UNAFOLD 3.8 at a temperature of 50 °C and a Na^+ concentration of 50 mM. The sequences were converted into MILP problem instances with MATLAB R2012a (Mathworks) (https://chilkotilab.pratt.duke.edu/codon-

scrambler). These instances were then solved using a MATLAB interface to CPLEX 12.5.1 (IBM) using a Dell Precision M4700 Laptop with 8 GB RAM and a quad-core i7-3820QM 2.70 GHz processor. iRBS were found by calculating the minimum hybridization energies of all arc-sequences with the *E. coli* anti-SD sequence, 5′-ACCTCCTTA-3′. ΔG values were calculated with UNAFOLD 3.8 at a temperature of 37 °C and a Na^+ concentration of 50 mM. For AB12, arcs with values lower than –8 kcal mol^{-1} were deleted from the graph. Other forbidden sequences included BsaI, NdeI, BseRI, BamHI, and XbaI recognition sites, as well as stretches of 6 Gs, 8 Cs, 8 As, and 8 Ts.

4. Gene libraries for the discovery of novel protein polymers

4.1 Introduction

Previously, a diverse set of IDPPs have been designed to explore the sequence space that drives LCST behavior at physiological temperature ranges (*6*). The motifs are characterized by a P-X_n-G motif where *n* varies from 0 to 4, and X residues and neighboring amino acids were chosen to target the hydropathy of tropoelastin. This heuristic strategy was largely successful, as the majority of sequences chosen had transition temperatures between 20°C and 80°C at polypeptide concentrations of 50 μM in PBS (pH 7.4).

Driven by this work, we wanted to determine the boundaries of sequence space of observable LCST behavior. This, of course, required us to relax the rules of hydropathy in order to obtain IDPPs with and without observable LCST behavior. As an

initial survey, we used the Codon Scrambling strategy to synthesize potential encoders of LCST behavior containing repeating sequence motifs coding the amino acid pattern: PGXXXX, where X represents any amino acid and is Gly at least once in each motif. When X was chosen at random from all 20 amino acids, we were often faced with difficulties with protein purification. Target proteins were often expressed insolubly in inclusion bodies, which required His-tag purification under denaturing conditions (8M Urea) and mostly had poor yields. High-yield purification via inverse transition cycling (ITC) was also not possible due to the required presupposition that the protein exhibits LCST behavior. Overall, a "global" search in the sequence space of PGXXXX was difficult due to the bias of insoluble sequences whose polymers were also difficult to purify. We surmised that when residues were chosen at random, a large number of sequences were simply too hydrophobic. Thus, despite the efficiency of the codon scrambling algorithm, a larger library of genes was required in order to screen out the large number of insoluble proteins from the sequence space. High throughput gene synthesis via codon scrambling at the scale of hundreds to thousands of genes can still be cost-prohibitive, at a current rate of approximately 0.1 USD per base (*40*).

Guided by our preliminary work, we developed a pooled gene-library based approach based on novel synthesis technologies to increase the number of potential encoders of LCST behavior, improving the rate of discovery of a broad "global" search of sequence space, with coverage beyond the "local" limits of bioinspired design. Following a global search, the Codon Scrambling approach synthesize polymers for

"local" search, by perturbing sequence content and length and acquiring cloud point data. With the limitations of synthesis of repetitive IDPPs overcome, the resulting cloud point data acquired can serve to establish a detailed and predictive insight on the rules that govern phase behavior.

From a gene synthesis perspective, this would allow for an unprecedentedly diverse synthetic gene library from oligonucleotides as starting materials. Overall, our gene library-based pipeline: (1) assembles large, complex gene libraries from a library of oligonucleotides, (2) transforms them in *Escherichia coli* for pooled expression and solubility screening on agar medium, and (3) individually characterizes IDPP candidates in a microwell format for phase diagram data. The following chapter contains work which will be published in a peer-reviewed journal.

4.2 Oligonucleotide libraries as materials for gene synthesis

Since their inception in 1995, microarrays have dramatically revolutionized genomics, with its massive parallelism and automation. Microarrays are 2D solid-phase arrays used to assay or screen biological materials like proteins or cells. Oligonucleotide arrays can be used for a variety of array designs, including gene expression screens, SNP genotyping, Comparative Genomic Hybridization (CGH), tiling, ChIP-on-chip, microRNA, resequencing, and aptamer screening (*84*). More relevantly, oligonucleotide microarrays have offered significant reductions in gene synthesis costs due to their massively parallel densities and throughput. For example, scaling down reagent volumes in resin-based synthesis is limited to decreasing the diameter of capillaries. Simply eliminating the rinsing of lines in instrumentation makes at least a 10-fold

reduction in cost (*85*). This is because acetonitrile is often used in high volumes for rinsing and dissolving reagents, making it one of the most expensive bulk reagents (*86*). High density arrays can offer 10^4-10^5 unique oligonucleotides, and can reduce costs by at least an order of magnitude (*43*, *87*). The variety in technologies and techniques for microarray-based oligonucleotide synthesis offer reductions in cost and advances in throughput, a major advantage for the Codon Scrambling technology. The technologies that have existed in the market include ink-jet printing (Twist Bioscience, Agilent, Protogene), photosensitive 5′ deprotection (Nimblegen, Affymetrix, Flexgen), photo-generated acid deprotection (Atactic/Xeotron/Invitrogen, LC Sciences), and electrolytic acid/base arrays (Oxamer, Combimatrix/Customarray).

4.2 RCA-based pooled polymer library synthesis

For our gene discovery phase, we developed a novel technology for the pooled, "one-pot" synthesis of repetitive gene libraries from microarray-based oligonucleotides on the order of one gene per oligonucleotide molecule. With the availability of microarray-based oligonucleotide pools available to researchers which consist of hundreds to thousands of oligonucleotides, this allows for the one-pot synthesis of hundreds to thousands of genes. This approach aims to produce the entire library in a single cloning step, leaving the primary costs to sequencing (~$6 per gene) as well as the initial purchase of oligonucleotides (~$1k per library) (Table 5).

Table 5: Approximate costs for gene library synthesis

	Codon scrambling strategy	Pooled library strategy
Cost for 50 genes	~$2500	~$2500
Cost for 500 genes	~$25000	~$4000

As a substitute for traditional gene synthesis methods, we use multiply-primed rolling circle amplification (RCA) to synthesize repetitive genes (*88*). One of the main restrictions of repetitive genes is that they are inaccessible to gene synthesis or *in vitro* amplification by polymerase chain reaction (PCR) (*89*). This is because they include annealing steps in which single-stranded DNA molecules would anneal out of register with each other at multiple sites. In contrast, RCA avoids annealing and amplifies DNA isothermally. A circular oligonucleotide template is randomly primed and polymerization proceeds to creates tandem repeats of the unit-length oligonucleotides at 30°C. This is accomplished with the φ29 DNA polymerase, a highly processive and strand displacing polymerase which continues to displace and 'roll off' non-template strands without dissociating from the template. A cascade of random priming events generates high molecular weight double stranded DNA from each individual template simultaneously.

Partial digestion of RCA products with a single appropriate restriction enzyme would yield clone-able gene fragments of various sizes. These fragments can be size-selected via gel electrophoresis and seamlessly ligated directly into expression plasmids with compatible ends. Partial digestion can be achieved by partial methylation with 5-

methylcytosine (5mC) (*90*). 5mC blocks digestion by certain restriction enzymes. During RCA, 5mC can be randomly incorporated along the DNA strands by including standard dNTPs and 5-methyl-dCTP in the reaction.

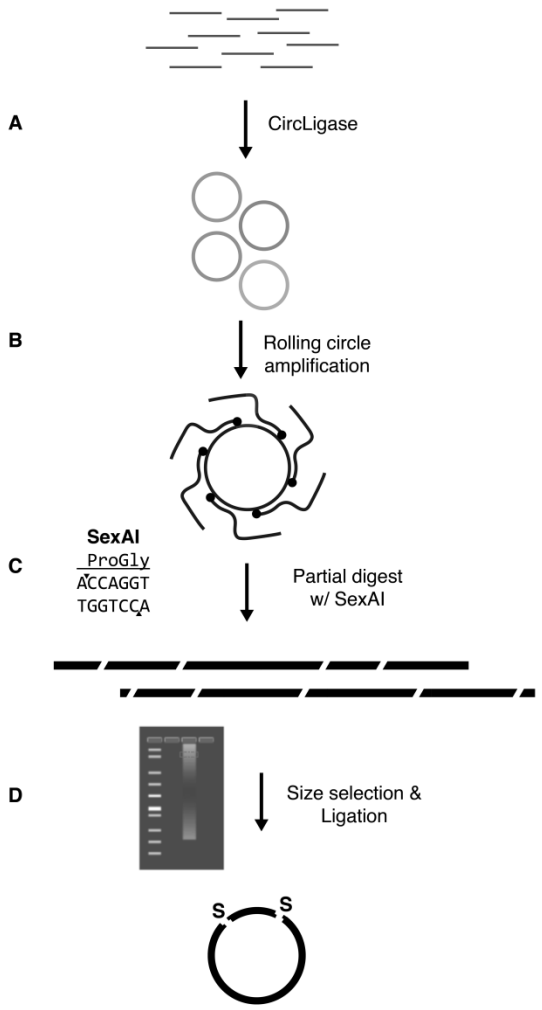

Figure 23: Generation of repetitive IDP libraries. a, As a starting material for peptide polymer gene-libraries, we will use LC Science's Oligomix. This product delivered as a pool of up to approximately 4000 customer-specified DNA oligonucleotide sequences in a single tube. Synthesis occurs on chips, in microreaction chambers connected to fluid microchannels. CircLigase then catalyzes intramolecular ligation of 5′ phosphorylated single stranded DNA molecules of an oligonucleotide library without producing concatemers. **b**, Circularized oligonucleotides serve as a template for RCA. In the reaction mixture, the substitution of dCTP with 10% 5-methyl-dCTP randomly incorporates 5mC into the RCA product. **c**, Partial restriction digestion of methylated RCA products generates gene fragments of various lengths) Gene fragments are

separated on a gel and a narrow range of lengths are isolated. The resulting mixture of DNA is then seamlessly and directionally ligated into expression plasmids using T4 DNA ligase.

We chose each of our oligonucleotide library's 72 bp protein coding sequences, which codes four motif repeats, to include at least one restriction enzyme recognition site. Pro/Gly amino acid pairs have been identified as structure-breaking residues and are ubiquitous in naturally occurring IDPs that exhibit temperature-triggered phase transitions. We therefore limited our search to repeating motifs containing the Pro-Gly amino acid pattern, and subsequently chose the SexAI restriction enzyme, whose recognition site overlaps with the Pro-Gly coding sequence. SexAI is methylation-sensitive and its 5 base pair recognition site is also non-palindromic, allowing for directional cloning of the gene fragments.

By using pooled oligonucleotide synthesis, we have control over the sequence content of the gene library. A 72-nucleotide oligonucleotide sequence encodes 4 repeats of a 6-amino acid motif. We selected a highly diverse set of 1020 oligonucleotides with four repeats of a sequence motif coding the amino acid pattern: $PGX_1X_2X_3X_4$, where X_n for n = 1, 2, 3, or 4, represents any amino acid except for proline and cysteine, and is glycine at least once in each motif (Table 10). With this pattern, proline content is ~15% and glycine content is ~30%, which is consistent with the composition of natural elastomeric domains (*7*). Cysteine is removed due to its propensity to form disulfide bridges. With 1020 sequences, all possible selections of amino acids for X_n are represented, ignoring permutations. Codons were chosen using the codon scrambling algorithm to remove inverted nucleotide repeats or hairpins in order to maximize

success of subsequent polymerase or ligase reaction steps (Table 11). We split the SexAI recognition site to the 5′ and 3′ ends of the oligonucleotide sequence, so that truncations, which are common in long oligonucleotide products, would contain deletions in the SexAI recognition site upon circular ligation. These products would not be digested by the SexAI enzyme and therefore would not be cloned into the expression vector.

Following partial digestion of RCA products, we were able to confirm the synthesis of gene fragments. Since we designed 72 nt oligonucleotides to contain exactly one recognition site, the RCA products contain recognition sites at every 72 bp. A ladder pattern on the gel reveals that partial digestion occurs at these sites.

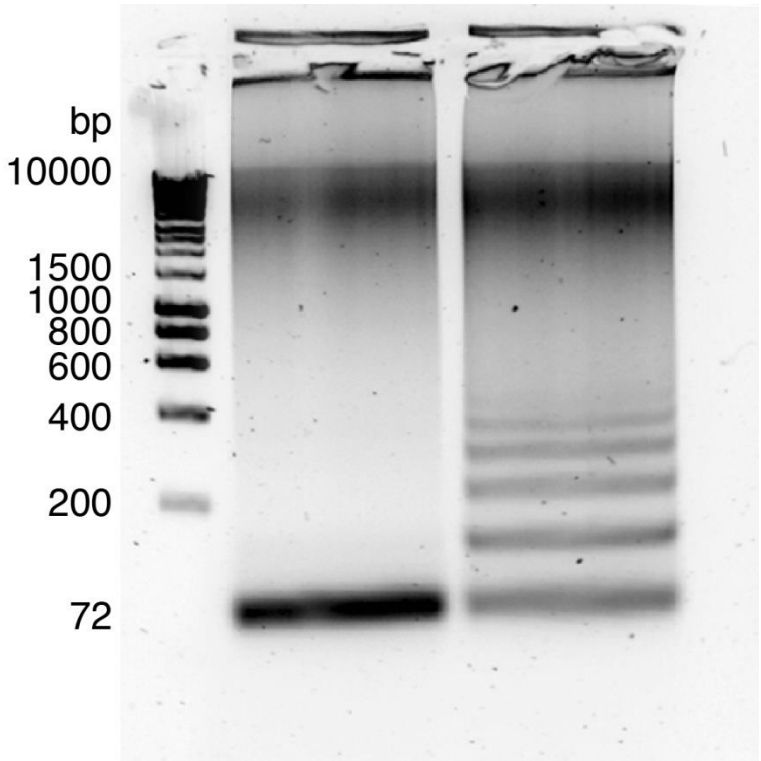

Figure 24: Restriction digestion of RCA products generated with 0% 5-methyl-dCTP (left) and 10% 5-methyl-dCTP (right). The gel's left lane shows a single band of 72 bp, the size of an oligonucleotide-length unit, indicating complete digestion. The gel's right

lane shows a distribution of lengths that are integer multiples of 72 bp, indicating partial digestion.

Oligonucleotide pools are often synthesized in small quantities, which is a concern because the pre-amplification of oligonucleotide pools is not trivial (*91*) and results a loss in diversity due to amplification bias. During optimization of RCA conditions, we found that compared to 72-nt circular oligonucleotides, 36-nt circular oligonucleotides were not efficient materials for library RCA, and that at least 125 fmol of starting oligonucleotide material was necessary for the gene synthesis procedure to generate a sufficient amount of DNA fragments.

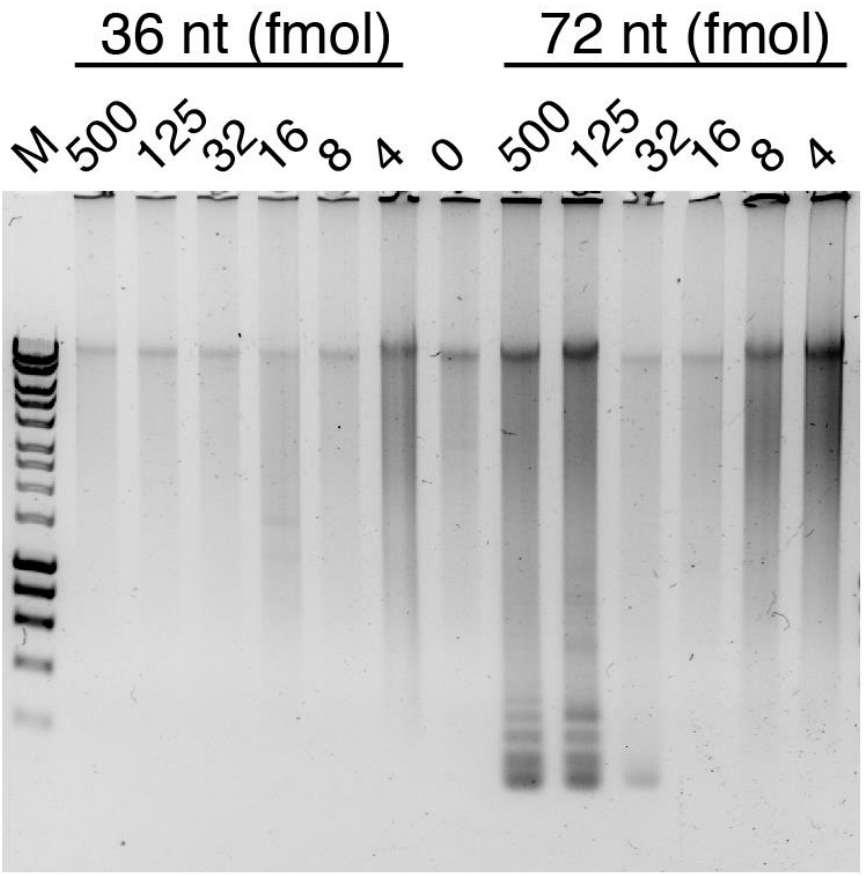

Figure 25: Partial digestion of RCA products from 36-nt oligonucleotide material (left) and 72-nt oligonucleotide material (right). A range of oligonucleotide concentrations indicate a lower limit for RCA-based gene library synthesis.

We isolated a gel slice containing DNA of 500-700 bp in length, and purified the DNA for cloning. Prior to cloning, we modified a pET24a *E. coli* expression plasmid to serve as a cloning vector for the gene fragments. This plasmid was modified include coding regions flanking a seamless SexAI cloning site.

4.3 Next generation sequencing verifies library diversity

Sequence patterns and diversity were identified with next generation sequencing (NGS). We prepared the plasmid library for Illumina Miseq sequencing in a workflow

similar to the preparation of metagenomic analysis of the 16S ribosomal RNA gene (*92*). We used a 2 × 251 bp paired end flow cell in the Illumina Miseq next generation sequencing (NGS) system. After sequencing, we used the Burrows-wheeler aligner (BWA) for mapping sequences generated from Illumina Miseq against the 1020 reference sequences which were used for oligonucleotide synthesis. The sequence alignment via Burrows-wheeler aligner (BWA) returned a list of compact idiosyncratic gapped alignment reports (CIGAR) and reference sequences for each sequence returned by NGS. The CIGAR string indicates alignment information including sequence matches (=), mismatches (X), insertions (I), and deletions (D) (Table 12). Out of 2 million sequences returned by NGS, there were 3888 unique alignments based on reference oligonucleotides paired with CIGAR strings. Alignments were then filtered for sequences that were too short, aligned poorly, or contained nonsense mutations, large frameshifts, or less than 3 perfect repeats. 865 unique sequences remained, which included perfect alignments and sequences with substitutions or small frameshifted regions. These remaining sequences aligned with 419 of the original 1020 oligonucleotide references due to the introduction of sequence bias. An additional 180 unique sequences aligned with reverse complements of the oligonucleotide references, resulting from cloning of a fragment in the incorrect orientation, but still resulting in a proline and glycine rich amino acid sequence. Sequence bias and loss of diversity was likely introduced in multiple steps, most likely from the synthesis of the relatively long 72-nt oligonucleotide source material, RCA and PCR amplification, and competition during library amplification in *E. coli*. To combat this, we used short RCA times,

minimal PCR cycles, and growth of *E. coli* colonies on solid agar instead of liquid media (*93*). After optimization, the biodiversity-loss was acceptable for our downstream purposes, as the resulting sequences were diverse in amino acid composition (Figure 26).

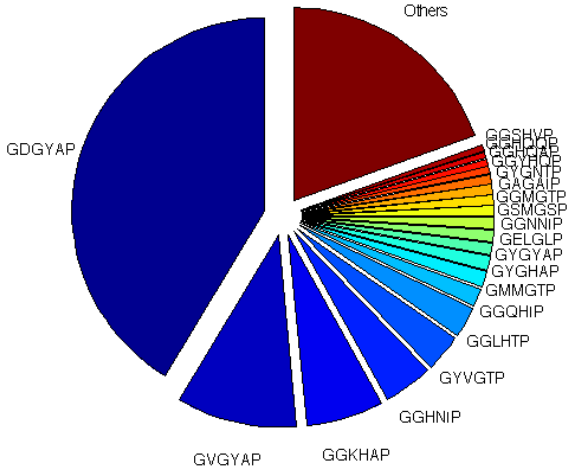

Figure 26: Illumina Miseq analysis of cloned library of plasmid DNA verifies the presence of at least 900 unique repetitive amino acid sequences.

4.4 Colony filtration screens for soluble expression

To rapidly identify soluble protein expression in *E. coli* while removing negative clones of the above categories, we have implemented the Colony Filtration (CoFi) blot method, a colony screening tool which has allowed us to identify genes encoding potentially soluble proteins (*94*). As discussed earlier, preliminary experiments indicated that our initial IDPP gene libraries encode a large number of insoluble proteins, presumably due to the large number of repetitive proteins with high hydrophobicity. The CoFi method positively screens for solubly expressed peptides at the colony level. It also serves two additional purposes. On the protein level, positive colonies are free of proteins that cannot be expressed due to metabolic burden, toxicity or decreased

translational rates due to poor codon usage. On the DNA level, positive colonies are free of recombinant plasmids with undesirable indel-induced frameshifts which prevent the translation of the His_6-tag or introduce premature stop codons. Both cases yield the absence of signal.

Briefly, colonies are transferred to a 45 μm membrane and are lysed in situ. Only soluble peptide passes through the membrane onto nitrocellulose, which is then used for His6-tag western blotting.

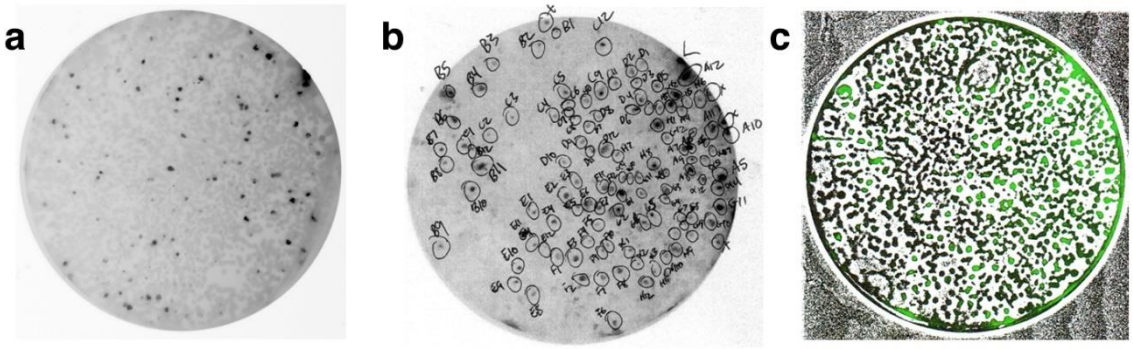

Figure 27: The CoFi method is used to identify and locate *E. coli* colonies expressing soluble proteins. a, Dark spots on a His_6-tag western blot indicate colonies expressing soluble proteins. **b,** These colonies were identified manually on a printed image and **c,** the thresholded image was used to align the identifications with the plate of colonies.

We further screen using PCR amplification directly from *E. coli* colonies with primers flanking the repetitive proteins to screen for empty vectors and short genes of undesired lengths which are not removed by the CoFi method. Short genes of undesired lengths were likely isolated with the longer genes during the size selection process due to aberrations in gel migration and/or cross-hybridization between different DNA molecules. For particularly dense plates, some colonies contained mixtures of genes with empty vectors or other genes, presumably due to cross-contamination of neighboring

colonies. These colonies are also identified with the colony-PCR screen and were removed from consideration for further processing (Figure 28).

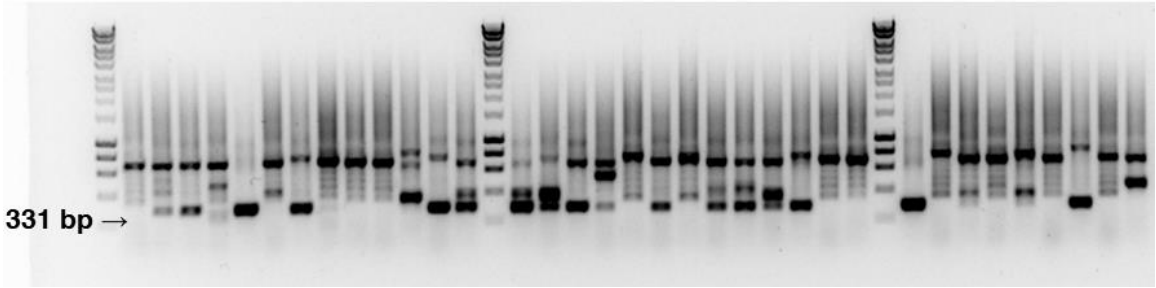

Figure 28: PCR directly from library-transformed *E. coli* colonies on dense plates visualized on agarose gel electrophoresis. Some colonies (positive) contained amplicons of expected and desired lengths, encoding 24-32 repeats of amino acid motifs (600-1000 bp). Each colony harbors an expression plasmid containing a repetitive gene. Other colonies (negative) contained amplicon of lengths expected to contain coding regions consisting of single-oligonucleotide lengths (331 bp).

Following colony screening, 572 colonies were sent to a sequencing vendor for direct DNA sequencing from *E. coli* cultures using sequencing primers that flanked the coding regions. Here, duplicates are removed, and the remaining colonies are labeled so that their sequences are tracked during the subsequent steps of protein expression (Table 6). Several sequences included missense mutations, which slightly changes the overall amino acid composition (Table 13).

Table 6: Summary of sanger sequencing results

Motif	Occurrences in NGS	Occurrences in sanger sequencing	Mutations
GLSGSP	6	3	1
GIRGTP	1	1	4
GGMGTP	8	1	1
GTGMEP	9	1	0
GGMGSP	11	1	2
GGTAAP	13	1	1
GNMGSP	20	1	0

GGSQIP	23	1	1
GTLGQP	38	1	0
GHMGTP	171	3	1
GGLSIP	59	1	0
GNIGTP	65	1	0
GGNYAP	66	1	2
GHHGTP	87	1	0
GGQHTP	194	2	0
GMGVAP	100	1	0
GEGNIP	223	2	1
GGHMIP	106	1	0
GKFGTP	114	1	0
GNGNVP	252	2	1
GGHQQP	605	4	0
GHIGVP	301	2	0
GTHGTP	145	1	0
GAGAIP	654	4	0
GIGQAP	188	1	0
GGMGIP	209	1	2
GTFGTP	219	1	0
GQSGLP	449	2	0
GAVGTP	242	1	0
GGHHAP	272	1	1
GSMGSP	609	2	0
GVIGIP	316	1	0
GGHQAP	345	1	0
GDNGSP	388	1	0
GGNNIP	414	1	1
GGSHVP	1189	3	0
GGLQQP	454	1	0
GGQHIP	1964	5	0
GHEGSP	487	1	0
GGMLAP	5076	13	0
GGHNIP	3826	5	0
GGLHTP	4024	5	0
GYGHAP	2763	3	1
GDGYAP	8094	11	1
GYGYAP	1450	1	0
GYMGKP	1558	1	0
GGFMQP	2975	2	0
GGKHAP	9432	7	2
GYVGTP	5938	3	1
GELGLP	4028	1	0
GGKHAP	12470	6	0
GVGYAP	5099	1	0

Following sequencing, we implemented a high-throughput protein production (HTPP) pipeline as an enabling technology for IDPP discovery. Recombinant proteins

are expressed from plasmid DNA from transformed *E. coli* BL21 suspension cultures, which offers a straightforward approach for HTPP (*95*). Following the protein production phase, we employ techniques to investigate the thermal phase transitions of these protein-based biomaterials. However, these techniques typically require high yields of purified protein. We adopted the Duetz Microflask system to cultivate cells with high growth rates and protein yields, at a throughput rate of about 48 IDPP candidates (*96*). Following the induction of the *E. coli* T7 protein expression, cells are harvested for purification. Soluble and insoluble fractions were separated from lysate and both were further characterized to further screen for solubility. The dot blot assay and polyacrylamide gel electrophoresis (PAGE) were both used to quantify the relative quantities of protein in soluble and insoluble fractions.

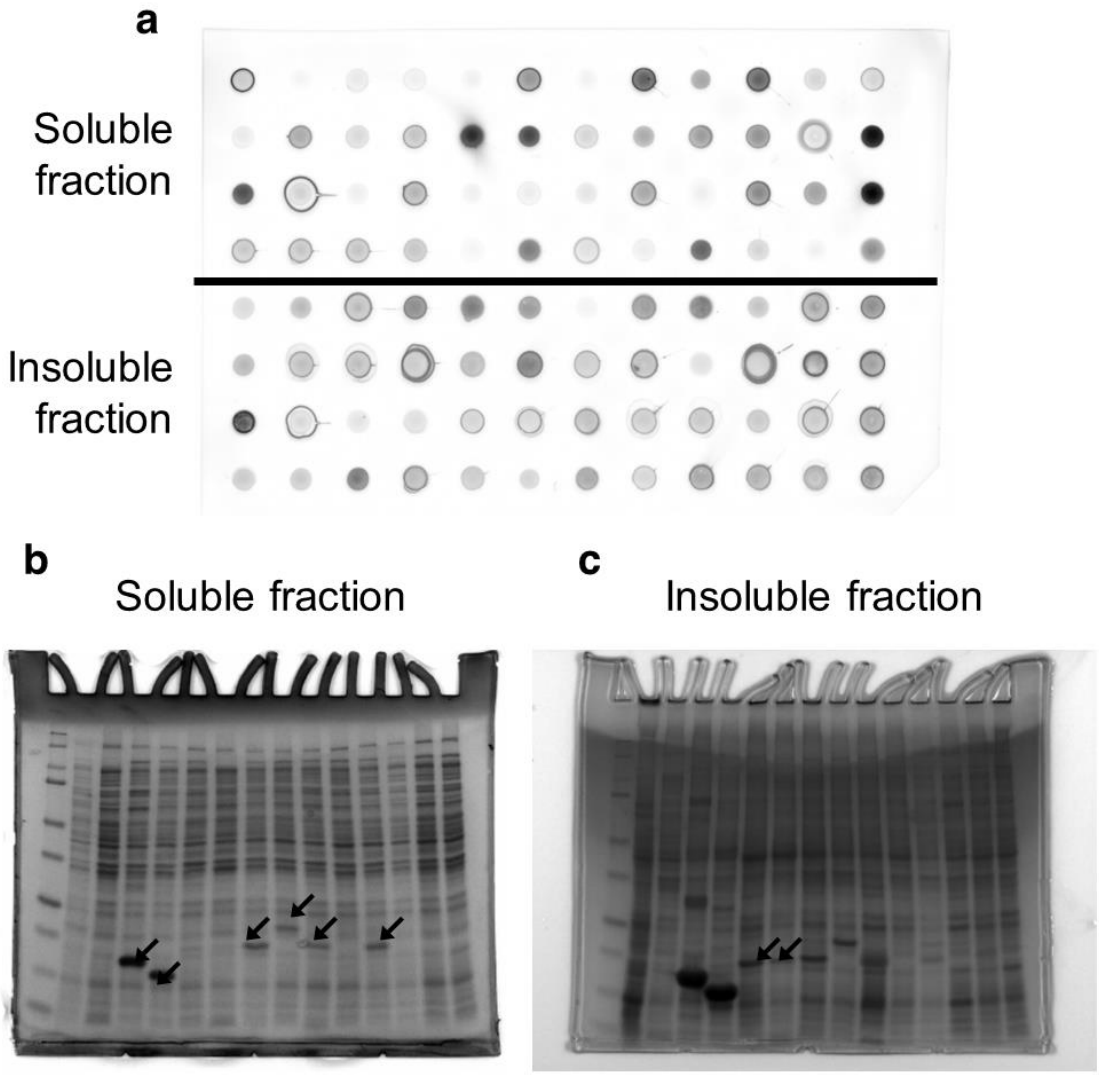

Figure 29: Examples of (a) dot blot and PAGE analysis of (b) soluble and (c) insoluble lysate fractions to determine target protein solubility. The distinct bands marked by arrows represent target proteins identified as soluble (**b**) or insoluble (**c**) by their relative presence between PAGE gels. 30 μg of total cell lysates were separated by centrifugation. Insoluble pellet was resuspended in 100 μLof 8M Urea. 10 μL of soluble fraction and 10 μL of insoluble fraction were subject to dot blot on a nitrocellulose membrane, which was then subject to His_6-tag antibody detection. 10 μL of soluble fraction and 10 μL of insoluble fraction were separated by SDS-PAGE on a TGX precast gel and then stained with EZBlue gel staining reagent (Sigma).

Cells are purified in one of two ways. For most proteins we used His_6-tag affinity purification with Ni-NTA spin columns (*97*, *98*) to directly purify recombinant His6-

tagged proteins in a rapid fashion. To purify proteins that appear to exhibit lower critical solution temperature (LCST), we modified the inverse transition cycling (ITC) protocol to process miniaturized volumes of cell cultures in a medium to high throughput. ITC uses the phase transition of temperature-responsive IDPPs to separate from *E. coli* contaminants with high yield (*99*).

4.5 Characterization of soluble proteins

We selected a subset of purified soluble proteins, and used circular dichroism to predict secondary structure.

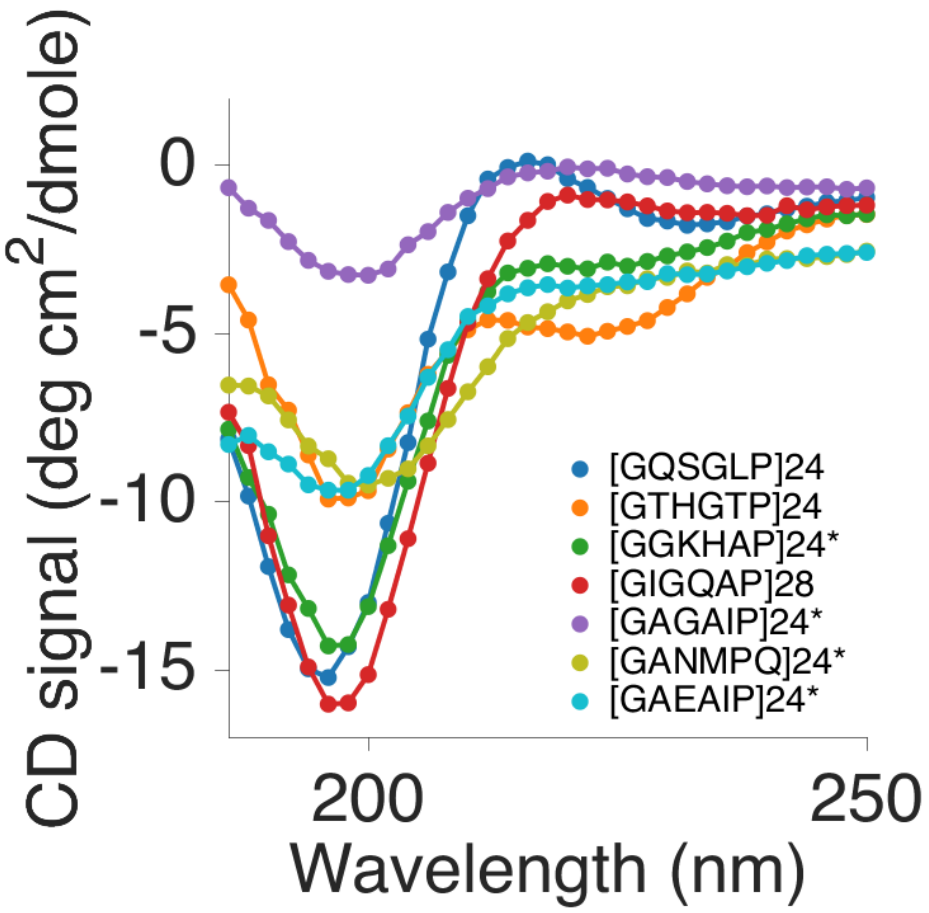

Figure 30: Protein polymers lack ordered secondary structures as shown by the CD spectra at 25 °C of a subset of soluble protein polymers. The shape of CD spectra are

characteristic of IDPs. [GGKHAP]24*, [GAGAIP]24*, [GANMPQ]24*, and [GAEAIP]24* have missense mutations and denote [GGKHAPGGKHAPGGKHAPDDKHAP]6, [GAGAIPGAGAIPGAGAIPVAGAIP]6, [GANMPQGASIPPGANIPPGASIPP]6 and [GAGAIPGAEAIPGAGAIPGAGAIP]6 respectively.

We identified a limited yet diverse set of soluble IDPPs that also exhibited phase transitions at physiologically relevant temperatures. These transitions are shown in Figure 31.

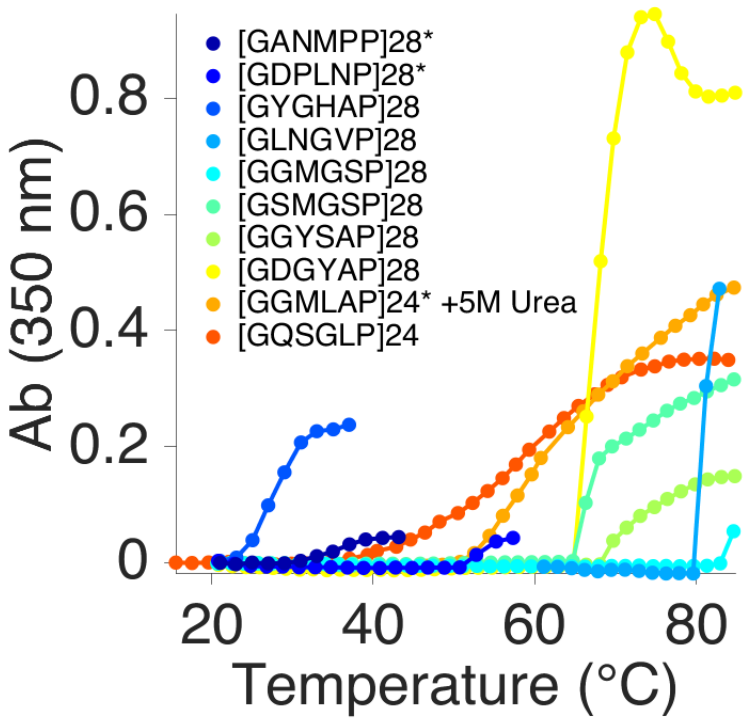

Figure 31: Turbidimetry data for protein polymers that exhibit LCST behavior between 15°C and 85°C in PBS, or PBS plus Urea for [GGMLAP]24*. [GANMPP]28*, [GDPLNP]28*, and [GGMLAP]24* have missense mutations and denote [GANMPPGASIPPGANIPPGASIPP]7, [GEPERPGLPLNPGDPLNPGLPDKP]7, and [GGMLAPGGMLAPGGMLAPGGMFAP]6 respectively. Polypeptide concentrations varied depending on purification yield.

4.6 Materials and methods

4.6.1 Gene library synthesis using RCA

A tube of OligoMix® was ordered from LC Sciences, containing approximately 3.5 pmol of 1020 unique specified 72-nt sequences. A 10μL PNK reaction was prepared containing 1 pmol 72-nt oligonucleotide mix, 1 X T4 Ligase buffer, and 10 U T4 PNK. The reaction was incubated at 37°C for 30 minutes and inactivated at 65°C for 20 minutes. A 20 μL CircLigase II reaction was prepared containing PNK-treated oligonucleotide mix, 200 U CircLigase II, 1X CircLigase II Buffer, and 2.5 mM $MnCl_2$. The reaction was incubated at 60°C for 2 hr and inactivated at 80°C for 10 min. To remove un-circularized ssDNA, a 12 μL Exonuclease reaction was prepared containing 10 μL CircLigase II product, 20 U Exonuclease I, and 200 U Exonuclease III. The reaction was incubated at 37°C for 1 hr and inactivated at 80°C for 15 min. A 36 μL annealing reaction was prepared containing 2 nmol Exo-resistant random primer, 1.1X RephiPHI buffer, and 2 μL Exonuclease reaction product. The annealing reaction was incubated in a thermal cycler with a slow temperature ramp from 95°C to 4°C, holding for 1 min every at every 2.3°C increment. A 40 μL RCA reaction was prepared containing 36 μL annealing product, 50 nmol dNTPs mixture where 25% of the dCTP was replaced with methyl-dCTP, 200 U RepliPHI polymerase. The RCA reaction was incubated at 30°C for 4 hr, and inactivated at 65°C for 20 minutes. The resulting DNA was then purified via isopropanol precipitation. A 25 μL digestion reaction was prepared containing 5 U SexAI, 2 μL of purified DNA, and 1X NEB Cutsmart buffer. The reaction was incubated

overnight at 37°C. The reaction was visualized on a 1% low melting point agarose gel and a 400-700bp band was gel-purified using the QIAquick gel extraction kit (Qiagen).

SexAI recognition sites (ACCTGGT), and the His_6-tag coding region were added to a modified pET-24a(+) vector using PCR. The construction of this modified vector from pET-24a(+) (Novagen) has been previously documented(*56*). 2.5 ng of the plasmid was linearized and amplified in a 50 μL PCR reaction containing 250 μM of each dNTP, 1X Herculase II PCR buffer, 0.25 μM each of universal forward and reverse primers and 0.5 μL Herculase II Fusion polymerase. The reaction was incubated at 95 °C for 2 min, followed by 5 cycles at 95 °C for 20 s, 56 °C for 20 s and 72 °C for 2 min 40 s, followed by 25 cycles at 95 °C for 20 s and 72 °C for 3 min, and a final step at 72 °C for 3 min. The linearized vector was gel-purified from a 1% low melting point agarose gel using the QIAquick gel extraction kit (Qiagen). A 90 μL digestion reaction was prepared containing 50 U SexAI, 5 μg of purified DNA, 3 U rSAP, and 1X Cutsmart buffer. The reaction was incubated overnight at 37°C. The resulting reaction was electro-eluted from a 1% low melting point agarose gel using a BioRad Model 422 Electro-Eluter.

A 10μL ElectroLigase reaction was prepared containing 1 μL ElectroLigase, 1 X ElectroLigase reaction buffer, 20 ng ~6 kbp linearized vector and 10 ng of ~0.5 kbp library insert. The ElectroLigase reaction was incubated at room temperature (25°C) for 30 min and inactivated at 65°C for 15 minutes. The resulting mixture diluted 1:3 with distilled water and 1 μL was transformed into 20 μL ElectroMAX DH10B electrocompetent *E. coli* cells ($\geq 10^{10}$ cfu/μg) using the BioRad MicroPulser Electroporator.

4×100 μL of SOC outgrowth was spread on 4 100mm diameter LB-kanamycin agar plates. Following overnight growth at room temperature, the resulting colonies from each plate were covered with 1 mL of LB-kanamycin and gently scraped off with cell platers. The cell suspensions were removed from the plates with a pipettor and combined in a single 15 mL conical tube. Plasmids were prepared from the cell suspensions using the QIAprep spin miniprep kit (Qiagen).

4.6.2 NGS analysis of gene library

Illumina MiSeq was used to analyze the sequence content of the gene library prior to screening. Illumina overhang adapters were added by PCR using region-specific primers surrounding the coding region. 12.5 ng of plasmid DNA was linearized and amplified in a 25 μL reaction containing 0.2 μM of each of the forward and reverse primers and 1 X KAPA Hifi HotStart ReadyMix. The reaction was incubated at 95 °C for 3 min, followed by 8 cycles at 98 °C for 30 s, 55 °C for 30 s and 72 °C for 30 s, and a final step at 72 °C for 5 min. 6 reactions were pooled together and Ampure XP bead cleanup was used to remove fragments greater than 500 bp. Next, indices and Illumina sequence adapters were added using the Nextera XT Index Kit using the same PCR reaction conditions as above. During Miseq sequencing, 20% PhiX genome was added to counter-balance the relatively low complexity and diversity of the gene libraries compared to genome libraries.

4.6.3 Colony screening for solubility and expression

One ng of miniprep plasmid was transformed into 20 μL Acella™ electrocompetent BL21(DE3) *E. coli* cells using the BioRad Micropulser Electroporator.

The SOC outgrowth was divided into 10 LB-kanamycin plates per transformation, resulting in about 500 colonies per plate after overnight growth at 37°C. The plates were replica plated twice and grown overnight at 37°C. CoFi blotting was performed on one set of replica plates (*94*). Briefly, colonies were transferred to a Durapore membrane. The Durapore membrane was transferred, colonies-side up, to an induction plate containing kanamycin and 200 μM IPTG for 4 hr at 37°C. The Durapore membrane was then transferred over a filter sandwich consisting of a nitrocellulose membrane and a Whatman 3MM paper drenched in B-PER lysis buffer containing 5 U/mL DNase I and 100μg/mL lysozyme. The filter sandwich was incubated for 30 min at room temperature. The blot was then freeze-thawed three times at -80°C for 10 min and then at 37°C for 10 min. The nitrocellulose membrane was then subject to antibody detection via standard Western blot analysis protocol, using a 1:4000 dilution of DyLight 650 conjugated 6x-His epitope tag antibody (Life Technologies, cat. no. MA1-21315-D650). The resulting blot was imaged using a Typhoon 9410 (GE Healthcare).

Image thresholding was performed in ImageJ to identify colonies containing soluble protein. The resulting binary image was printed and the second replica plates were placed over the printed image and aligned. The binary image was then used to locate and pick colonies. Picked colonies were grown in 96 deep-well polypropylene microplates overnight at 37 °C and 200 rpm. Clones were screened for frameshifts, empty vectors, and duplicates with direct sequencing from colonies using T7 sequencing primers. Unique clones were saved for protein purification.

4.6.4 Protein expression and purification

As a validation step, a selection of genes was back-translated from amino acid sequences using codon scrambling (*40*) and resynthesized via high throughput DNA synthesis by vendors, Gen9 Inc., MA, USA and Twist Bioscience Inc., CA, USA (Table 15). A C-terminal trailer encoding for GW were incorporated into the genes, and all genes were cloned into modified pET-24a(+) plasmids. The resulting plasmids were transformed into BL21 competent *E. coli* cells. These sequences were purified in parallel with pooled library constructs to validate downstream characterizations.

The Duetz Microflask system was used to grow screened clones in 24 deep-well microplates holding 14mL round-bottom tubes containing 2 mL of Overnight Express TB (EMD Millipore) plus 100 μg mL^{-1} kanamycin (*96*). A clamp system and sandwich covers were used to allow for high-frequency orbital shaking and high oxygen transfer rates of 1-2 culture volumes per minute without well-to-well contamination. After overnight growth at 30°C and 1200 rpm on a VWR incubating mini shaker, chemical lysis was performed by resuspending cell pellet in 900uL of B-PER containing 3.6 U DNase I and 90μg lysozyme at room temperature. Three freeze-thaws were performed at -80°C for 15 minutes and room temperature for 15 minutes. Afterwards, target proteins were purified from lysate by His6-tag purification or inverse transition cycling (ITC).

His6-tag purification was performed with 2 mL spin columns containing 1 mL HisPur TM Ni-NTA resin under denaturing conditions (8M Urea). Urea was removed

by overnight dialysis in Thermo Scientific™ Slide-A-Lyzer™ Dialysis Cassettes in 10L ddH2O at 4°C.

ITC was performed in 24-deepwell polypropylene plates. 5M NaCl was titrated to induce phase transition. The pellet was centrifuged at 2800g for 15 minutes at 45°C. The supernatant was removed and the protein pellet was then resuspended in 300 μL PBS at 4 °C. 10 μL of PEI was added to the lysate. The resuspension was heated to 70 °C for 10 minutes and then cooled to 4 °C on ice for 10 minutes. PEI-precipitated DNA and natural proteins with irreversible temperature-dependent aggregation were removed by centrifugation at 2800g for 15 minutes at 4°C.

For insoluble proteins, the above protocol was slightly modified. The lysate was first centrifuged at 2800g for 15 minutes at 4 °C. The pellet was resuspended in 300 μL of 8M Urea. The clear solution was then diluted with PBS until precipitate started forming, around 1-2M Urea. Then, the solution was centrifuged at 2800g for 15 minutes at 45 °C. Finally, the pellet was resuspended in around 3M Urea, increasing the concentration as necessary.

4.6.5 Cloud point characterization

To characterize the cloud point transition temperatures of IDPPs, the 350 nm optical densities of 25 μM IDPP solutions in PBS were monitored with a Cary 300 UV-Vis spectrophotometer (Agilent Technologies) as a function of solution temperature. The SYPRO Orange assay was also used to characterize cloud point transition temperatures. 5000X SYPRO Orange was diluted to 200X working stock using PBS. 20 μL of 25 μM IDPP solutions containing PBS and 1X SYPRO Orange were prepared in low profile

MicroAmp optical qPCR tube strips. The StepOnePlus™ Real-Time PCR instrument was used to monitor fluorescence as a function of solution temperature using the ROX dye settings.

5. Additional tools for peptide polymer production

5.1 SYPRO orange assay for the high throughput determination of phase behavior

Because transition temperature is dependent on a number of parameters including sequence content, length, and concentration, model development requires a large number of experimentally measured cloud points. To obtain accurate thermal phase transition curves, Chilkoti and coworkers have used turbidity absorbance profiles, from which LCST behavior can be inferred. The cloud point is defined as the temperature upon the onset of turbidity during a slow and constant (~0.1 °C/min) temperature increase. The onset can be determined in a number of ways, including: the temperature at which the tangent of the maximum slope crosses the temperature axis, visual inspection of the initial deviation from baseline, the temperature at 50% of the maximum absorbance, the initial deviation of the first derivative, and the temperature at the maximum slope (*1*, *100*, *101*). For the cloud points in this work, the first, "maximum slope" method was used. An alternative, isothermal method to trigger the phase transition involves changing the ionic strength of the solution via the type and concentration of cosolutes: the addition of sufficient levels of a kosmotropic salt such as sodium citrate or a chaotropic salt such as urea can depress or elevate LCST phase

transition temperature into observable ranges (0 – 100° C) relative to that of water, respectively by changing the chemical potential of the solvent (*102*).

Unfortunately, traditional spectrophometric machines can thermally probe less than 10 samples at a time, which limits the throughput of LCST determination. Currently, high throughput screening methods do not exist to investigate the phase transitions of protein-based biomaterials. Optical techniques such as turbidimetry (*103*), circular dichroism (CD) (*104*), infrared spectroscopy (IR) (*105*, *106*) and dynamic light scattering (DLS) (*107*) are routinely used to study protein thermodynamics. However, these techniques typically require high concentrations (>1 mg/ml) of purified protein and probe at most, several samples at a time. Here, we adapt the thermal shift assay, used to measure protein denaturation, for the phase transition of IDPPs. The thermal shift assay measures protein denaturation by using environmentally sensitive dyes such as SYPRO Orange, whose fluorescence increases with the increasing hydrophobicity of the environment (*108*).

Coil to globule transition of IDPPs mark the initial appearance of globule surface tension, as globules behave as spherical droplets whose surface tension originates from the contact between solvent and monomers at the surface. The rapid appearance of globules in solution increases the overall hydrophobicity of the environment. Therefore, the fluorescence mechanism of SYPRO Orange is well matched to phase transition of LCST polymers (Figure 32).

Using fluorescence to monitor phase transition has a number of advantages. High throughput 96-well Real-Time PCR machines such as the Applied Biosystems®

StepOnePlus™ Real-Time PCR instrument, are well equipped for protein thermal shift experiments; the peltier-based system uses LEDs as the excitation source, and filtered photodiodes for emission. Additionally, lower volumes of purified protein are necessary compared to turbidimetry due to the high fluorescence sensitivity and fast thermal ramp rates of the instrument.

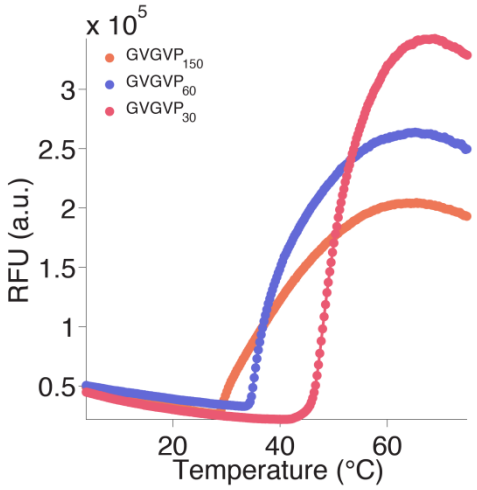

Figure 32: SYPRO orange assay with various lengths of elastin like polypeptides (ELPs). Thermal profiles were acquired using a StepOnePlus™ Real-Time PCR instrument at a range of temperatures with a total volume of 20 μL of approximately 25 μM purified protein in 1X PBS. A working concentration of 10X SYPRO orange was used.

For peptide polymers expressed purified from gene libraries, the SYPRO orange assay are used to rapidly assess the presence of a cloud point for many sequence compositions, concentrations, and lengths. Here, we show that the SYPRO orange assay effectively detects the phase transition of a diverse set of ELPs with varying amino acid fractions of alanine (Figure 33a) (*20*). All transition temperatures determined with the first derivative method are within 1.5 °C of previously determined transition temperatures from turbidimetry (Figure 33b). We also show detection of thermal phase

transition of peptide-polymer based micelles with concentrations as low as 1 μM (Figure 33c). For the same peptide-polymers, turbidimetry signals are dominated by noise for concentrations lower than 10 μM.

The SYPRO orange assay effectively detects the phase transition of the LCST polymers using a low volume (20 μL) and concentration (25 μM) of purified protein, which can be used to identify cloud points for IDPPs expressed and purified from gene libraries.

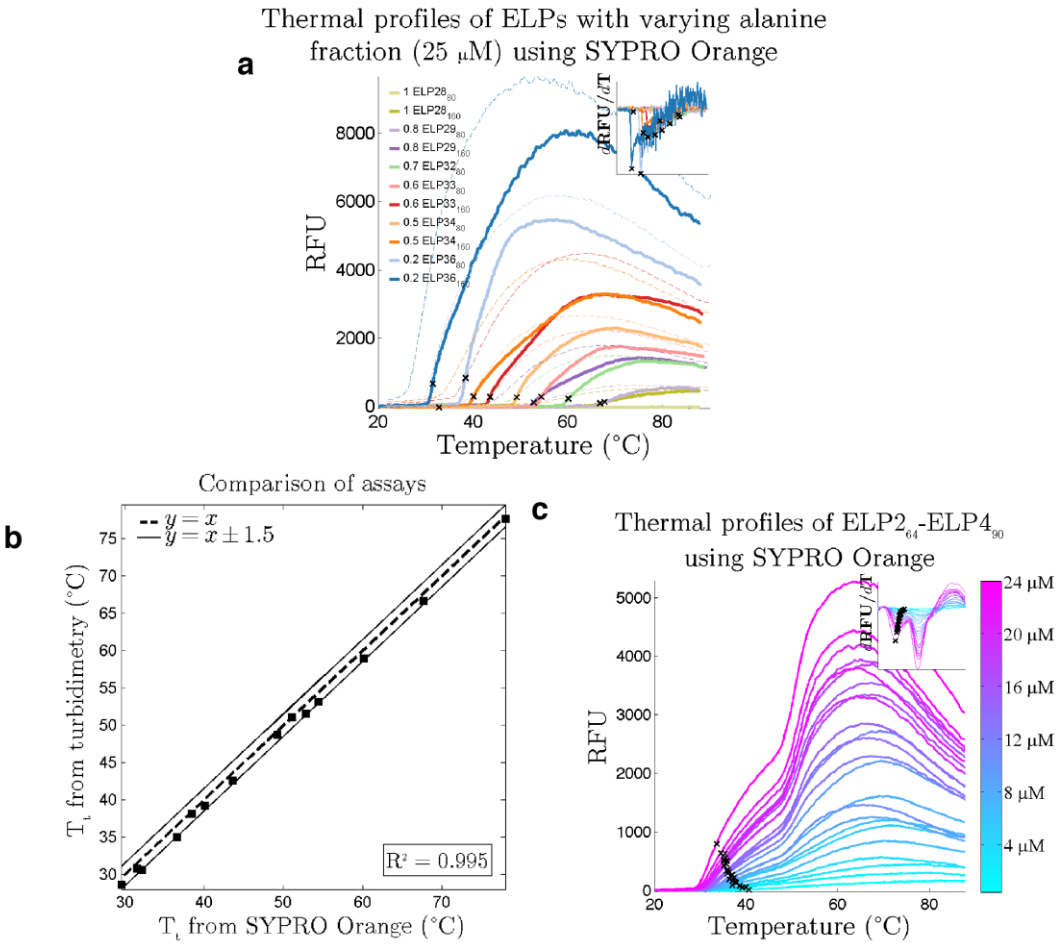

Figure 33: SYPRO orange assay with ELPs. a, Thermal profiles for 20 μL of 25 μM purified ELPs in 1X PBS and 10X SYPRO orange. Minimum peaks for dRFU/dT were used to determine the characteristic transition temperature. **b**, Linear correlation between turbidimetry and SYPRO orange determined transition temperatures. **c**, Thermal profiles of varying concentrations of ELP micelles in 1X PBS and 10X SYPRO

orange. The critical micelle temperatures (CMT) were calculated using the "maximum slope" method.

5.2 Unstructured non-repetitive polypeptides have LCST behavior

The modular organization of natural proteins suggest that they did not arise from completely random sequences, but developed from primordial genetic units from molecular mechanisms such as exon shuffling (*109*). However, motifs involved with structural function are often strongly influenced by their context within the protein. With the incomplete knowledge of structure-function relationships of proteins, *in vitro* combinatorial approaches are still necessary to transplant the structural-based functionality of motifs identified from natural proteins into artificial proteins (*110*).

Natural intrinsically disordered proteins such as tropoelastin are also interspersed with non-repetitive units. However, recent work suggests that LCST behavior is much less context-limited than previously thought. LCST behavior is retained in ELPs with interspersed regions of structural units composed of polyalanine helices (*111*). Also, relatively simple heuristics can be used to tune the sequence of repetitive motifs and still retain LCST behavior (*6*). While the sequence complexity of tropoelastin is quite high compared to synthetic proteins such as ELPs, the complexity is mostly attributed to sequence ordering rather than composition. Bioinformatics analysis of tropoelastin have identified a limited abundance of amino acids besides proline and glycine which include, in order of abundance: valine (V), alanine (A), leucine (L), lysine (K), threonine (T), isoleucine (I), tyrosine (Y), serine (S), and phenylalanine (F) (*6*).

From a scientific perspective, identifying LCST behaviour in nonrepetitive synthetic proteins would help bridge the gap between natural stimuli-responsive proteins and artificial ones. This would potentially extend the current understanding of phase behaviour and hydrophobicity of model proteins containing periodic repeats to more general sequence space of disordered proteins. From a practical standpoint, due to their higher complexity, much longer sequences can be synthesized by gene synthesis vendors than their repetitive counterparts without being rejected by their sequence complexity filters. A more apparent application of increased sequence complexity is the generation of a higher diversity of potential sequences with potential biological functionality, as sequences can be rearranged or mutated in more ways compared to modular gene fusions with functional proteins (*112*).

To expand sequence diversity, we have identified unstructured polypeptides having no discernible repeat sequence that also exhibit LCST phase behavior. We designed an algorithm that promotes well-distributed Pro and Gly, while maintaining amino acid fractions of their repetitive counterpart sequence of the form: [PGGXXX]40. Each amino acid sequence comprises 1/6 proline (P) residues, 1/3 glycine (G) residues, and 1/2 X residues, where X is one or more amino acids selected from the group consisting of valine (V), alanine (A), leucine (L), lysine (K), threonine (T), isoleucine (I), tyrosine (Y), serine (S), and phenylalanine (F). Each of the selected amino acids for X occurs at equal frequencies to each other. The algorithm was used to identify sequences with the exact composition, but permuted ordering of existing repetitive ELPs known to exhibit LCST behavior. The generated sequences were sufficiently non-repetitive; they

were rejected if they contained at least one motif comprising 5 to 10 amino acids that (a) contained three contiguous identical amino acids, (b) occurred more than once within the sequence, or (c) contained at least 2 prolines (P) separated by zero or more amino acids that are not glycine (G). These constraints reject sequences with local clusters of identical amino acids, and therefore promote well distributed prolines (P) and glycines (G) which have together been identified as structure-breaking residues.

We use a constraint satisfaction algorithm to identify sequences that satisfy the above requirements. First, similar to the codon scrambling strategy, we utilize a De Bruijn subgraph (*49*) to sparsely represent all feasible peptide sequences. A *k*-mer De Bruijn subgraph of a 20 amino acid alphabet is a directed graph in which each vertex represents a substring consisting of *k* amino acids. Each arc is a substring consisting of *k* + 1 amino acids, and connects its prefix and suffix vertices; for example, arcs between vertices corresponding to 5 amino acids can be represented by 6 amino acids, as shown by the overlapping amino acid sequences below:

$$PGAVG \xrightarrow{PGAVGV} GAVGV\ , \qquad \textbf{(48)}$$

where $k = 5$. In this sense, a path traversal across multiple arcs represents a sliding window, where a window of $k + 1$ amino acids incrementally advances across a sequence. We can then construct a graph such that feasible paths represent satisfy the constraints. Arcs are removed if their amino acid sequence contains three contiguous identical amino acids, or contains at least 2 prolines separated by zero or more amino acids that are not glycine.

Table 7 is the MILP programming formulation of the constraint satisfaction problem. It uses the same notation as Table 1 of the codon scrambling problem, except that nucleotide subsequences are replaced with amino acid subsequences and w_{ij}^{m} denotes the number of occurrences of amino acid m in arc ij. Briefly, x_{ij} is equal to 1 if and only if arc $ij \in A$ is in the optimal path. Constraints (a and f) impose that arcs are connected and distinct, constraint (b) imposes that the path is acyclic, constraints (c and d) impose that the path is Hamiltonian (vertices do not occur more than once in a feasible path), and constraint (e) ensures that the total composition is fixed. Sequences satisfying the constraints were identified using CPLEX. The first sequence to be identified was used for gene synthesis.

Table 7: MILP formulation of the constraint satisfaction program

$$\sum_{j \in V} x_{ji} - \sum_{j \in V} x_{ij} = \begin{cases} -1 & i = s \\ 1 & i = t \\ 0 & \forall i \in V \setminus \{s,t\} \end{cases} \qquad \forall i \in V \qquad \text{(a)}$$

$$u_i - u_j + p x_{ij} \leq p - 1 \qquad \forall ij \in A, i \neq s, j \neq t \qquad \text{(b)}$$

$$\sum_{j \in V} x_{ij} \leq 1 \qquad \forall ij \in A, i \neq s, j \neq t \qquad \text{(c)}$$

$$\sum_{i \in V} x_{ij} \leq 1 \qquad \forall ij \in A, i \neq s, j \neq t \qquad \text{(d)}$$

$$\sum_{ij \in A} w_{ij}^{m} x_{ij} = \begin{cases} 0.2L & m = A \\ 2L & m = G \\ L & m = P \\ 1.8L & m = V \end{cases} \qquad \forall m \in M \qquad \text{(e)}$$

$$x_{ij} \in \mathbb{Z}_2 \qquad \forall ij \in A \qquad \text{(f)}$$

We first identified two well-characterized ELPs, with repeating subsequences of GAGVPGVGVP (A[0.5]) or GVGVPGVGVPGAGVPGVGVPGVGVP (A[0.2]) (*20*). In this notation, A[Z] refers to poly[GXGVP] where Z is the fraction of valine-to-alanine substitutions in the guest residue, X. We used the algorithm to generate the permuted or rearranged versions of 4 ELP sequences such that they were maximally non-repetitive: A[0.2] rep-200, which includes the A[0.2] subsequence repeated 8 times for a total of 200 amino acids, A[0.5] rep-200, which includes the A[0.5] subsequence repeated 20 times for a total of 200 amino acids, A[0.2] rep-400, which includes the A[0.2] subsequence repeated 16 times for a total of 400 amino acids, and A[0.5] rep-400, which includes the A[0.5] subsequence repeated 40 times for a total of 400 amino acids. The amino acid compositions of these sequences are depicted in Figure 34.

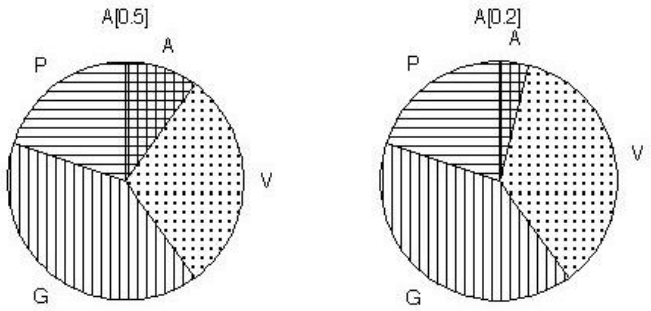

Figure 34: Amino acid compositions of the nonrepetitive and repetitive versions of the A[0.5] and A[0.2] sequences.

Following gene synthesis, these sequences have been expressed and purified from *E. coli*. Despite the absence of discernable repetitiveness, unstructured peptides exhibited lower critical solution temperature (LCST) behaviors that similar to their repetitive counterparts (Figure 35). Turbidity profiles, consisting of heating and cooling curves between about 0°C and about 80°C, for all constructs at 25 μM in PBS show that

non-repetitive versions of A[0.2] and A[0.5], retain reversible LCST phase behavior with transition temperatures less than 10°C different from their repetitive counterparts.

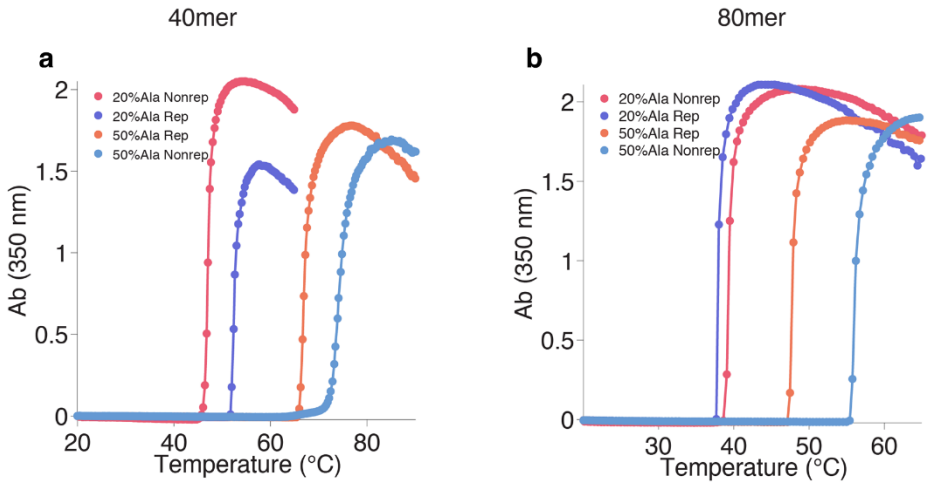

Figure 35: Characterization of repetitive and nonrepetitive versions of A[0.2] and A[0.5] comprising (a) 200 amino acids and (b) 400 amino acids. Turbidity profiles for all constructs at 25 μM in PBS show that non-repetitive versions of A[0.2] and A[0.5], and A[0.5] Nonrep retain LCST phase behavior with transition temperatures similar to their repetitive counterparts.

The algorithm was used to additionally generate a panel of 8 non-repetitive sequences, each comprising 240 amino acids with 1/6 being proline (P) and 1/3 being glycine (G). The remaining amino acids were various combinations of 9 different amino acids found in elastin (Table 14). These nonrepetitive sequences did not have existing repetitive counterparts. However, their amino acid compositions were chosen to target the hydropathy of tropoelastin (Figure 36). The notation adopts the form $X_1X_2\ldots X_n$, which refers to *n* amino acids, X_1, X_2, … X_n. which equally contribute to ½ of all amino acids in the total sequence, and proline and glycine contribute to the other half.

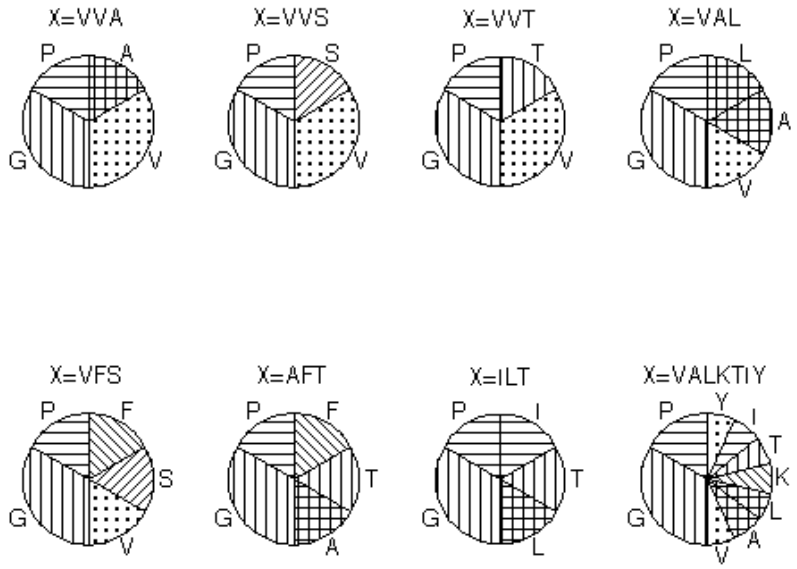

Figure 36: Amino acid compositions of exemplary non-repetitive unstructured polypeptides. Each non-repetitive unstructured polypeptide comprises 240 amino acids comprising 1/6 proline (P) residues, 1/3 glycine (G) residues, and 1/2 X residues, where X is one or more amino acids selected from the group consisting of valine (V), alanine (A), leucine (L), lysine (K), threonine (T), isoleucine (I), tyrosine (Y), serine (S), and phenylalanine (F). Each of the selected amino acids for X can occur at equal frequencies to each other.

After gene synthesis, expression and purification in *E. coli*, half of the sequences had observable LCST behavior at 25 μM in PBS. The other half did not have LCST behavior at physiological temperature ranges (0°C - 80°C), but did have observable LCST behavior upon addition of various concentrations of urea (Figure 37). These sequences had higher hydrophobicities, which likely explained the requirement of urea to elevate the transition temperature to observable ranges (*102*). Additionally, we attempted to express and purify sequences of compositions X = VVI, LKF, VVL, VIL, STY, VVF, VTY, AFS, IKY, AKY, and VVY using the same procedure and notation as those described in Figure 36. However, these sequences failed to overexpress, as they were not present in lysate in sufficient quantities, as determined by PAGE analysis. We

surmise that this may be due to the presence of large numbers of hydrophobic residues: isoleucine, leucine, tyrosine, and phenylalanine. Likewise, sequences of compositions X = VVK, VAT, and VALKT were expressed and purified, but did not transition due to hydrophilic residues, lysine and threonine.

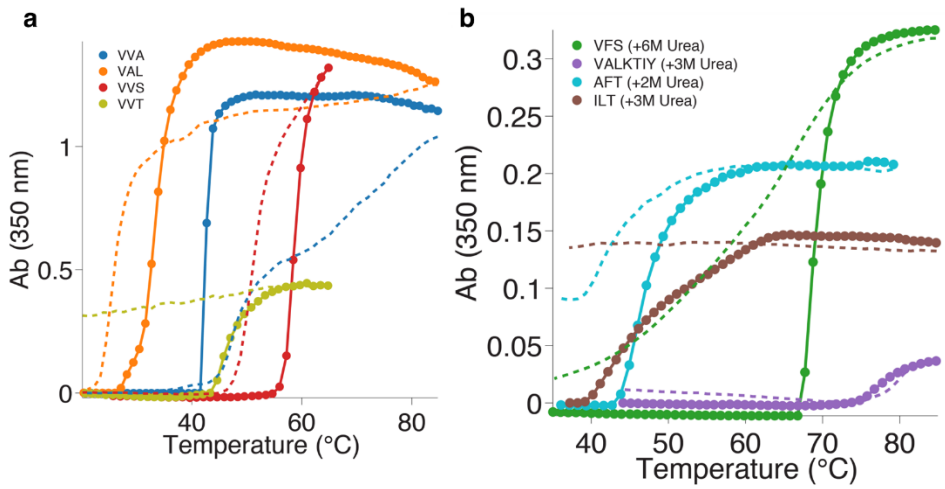

Figure 37: Transition temperature characterization of exemplary non-repetitive unstructured polypeptides with **a,** observable LCST behavior at 25 μM in PBS or **b,** observable LCST behavior upon addition of various concentrations of urea. Solid lines indicate heating curves and dotted lines indicate cooling curves.

As mentioned above, the unstructured polypeptides lack any discernable repeat motif of amino acids. The repetitiveness of the unstructured polypeptides may be characterized by its linguistic complexity score. Linguistic complexity score is defined by the total number of unique subsequences in a given sequence divided by the total number of unique subsequences possible for the same alphabet and window length (*113*).

The repetitiveness of each of the unstructured polypeptide sequences was quantified by calculating linguistic complexity scores. Linguistic complexity is defined by the total number of unique subsequences in a given sequence divided by the total number of

unique subsequences possible for the same alphabet and window. Scores were calculated using the protein analysis tool, CIDER (*114*). The window length was set equal to the total sequence length for each sequence. The final score is given as the product of the linguistic complexity score and total sequence length to account for sequence length. The resulting final scores for the repetitive polypeptide sequences in are reported in Table 1. All final scores for the non-repetitive polypeptide sequences were greater than 15.0.

Table 8: Characterization of the repetitiveness of repetitive and non-repetitive polypeptide sequences

Sequence	Product of length and linguistic complexity score
A[0.2] rep, 200 aa	7.9
A[0.5] rep, 200 aa	7.9
A[0.2] rep, 400 aa	8.0
A[0.5] rep, 400 aa	8.0
A[0.2] nonrep, 200 aa	28.7
A[0.5] nonrep, 200 aa	35.6
A[0.2] nonrep, 400 aa	32.8
A[0.5] nonrep, 400 aa	42.8
VVA	50.6
VVS	51.6
VVT	50.6
VAL	83.3
VFS	85.3
AFT	88.3
ILT	86.3
VALKTIY	147.8

5.3 Materials and methods

All constraint satisfaction problems were converted into MILP problem instances with MATLAB R2012a (Mathworks). These instances were then solved using a

MATLAB interface to CPLEX 12.5.1 (IBM) using a Dell Precision M4700 Laptop with 8 GB RAM and a quad-core i7-3820QM 2.70 GHz processor.

All nucleotide sequences, were back-translated from amino acid sequences using codon scrambling (Table 15) (*40*). A N-terminal leader sequence encoding MSKGPG and a trailer encoding for GWP were incorporated into the genes. All genes were synthesized by commercial synthesis and cloned into modified pET-24a(+) plasmids (Gen9 Inc., MA, USA and Twist Bioscience Inc., CA, USA). The resulting plasmids were transformed into BL21 competent *E. coli* cells.

Colonies were inoculated in 2-5 mL of Terrific Broth (TB) plus 50 μg/mL kanamycin and grown overnight at 37°C and 250 rpm. One milliliter of the starter cultures was inoculated in 1 L of TB plus 50 μg/mL kanamycin and grown for 6-7 h at 37°C and 250 r.p.m. Expression was induced by the addition of IPTG at a final concentration of 1 mM, and the cells were grown for an addition 24 h. The unstructured polypeptides were purified by inverse transition cycling (ITC) as previously described (*103*). To characterize the inverse transition temperature of unstructured polypeptides, the 350 nm optical densities of 25 μM peptide solutions were monitored with a Cary 300 ultraviolet–visible spectrophotometer (Agilent Technologies) as a function of solution temperature.

6. Validation of the phase transition model and future directions

In summary, we have built a general model of IDPPs exhibiting phase transition behavior in chapter 2, which can be extended to calculations of globule surface tension.

We have also generated a set of positive and negative data for LCST behavior. To achieve this, we have used a combination of two strategies intended to improve the rate of discovery of new stimuli responsive peptide polymers: (1) Our published Codon-scrambling algorithm enables the high throughput PCR-based gene synthesis of repetitive proteins by exploiting the codon redundancy of amino acids and finding the least-repetitive synonymous gene sequence (Chapter 3); (2) our novel pooled library approach uses rolling circle amplification (RCA) to create large complex gene libraries from libraries of oligonucleotides (Chapter 4). Using both technologies, we have generated ~60 new polypeptides with motifs of the form [PGGXXX]n, where X is a variable amino acid and n ranges from 24-28. Starting from a DNA library size of at least 1000 unique gene sequences (verified by next generation sequencing (NGS)) we screened for cloning artifacts such as short sequences, internal stop codons, and frameshifts, as well as for expression and to some extent, solubility. A subset of the screened sequences was purified and further characterized by turbidimetry for cloud-point transition temperature, and verified to be disordered and soluble with circular dichroism (CD) and polyacrylamide gel electrophoresis (PAGE).

To expand sequence diversity, we have also identified unstructured polypeptides having no discernible repeat sequence that also exhibit LCST phase behavior (Chapter 6). We designed an algorithm that promotes well-distributed Pro and Gly, while maintaining amino acid fractions of their repetitive counterpart sequence of the form: [PGGXXX]40. This algorithm ensures a maximally nonrepetitive sequence by ensuring that every subsequence of 5-10 amino acids are unique. We have found that

maximally nonrepetitive sequences and their repetitive counterparts maintain similar LCST behavior. This category of sequences further adds to the sequence flexibility and data set that can be used to model LCST behavior.

We have provided a LCST data set which offers a starting point to further understand our models to predict LCST behavior. We have noticed that average residue Urry hydropathy, a descriptor based on linear models of the critical transition temperatures of ELP with varying guest residues, can be used to separate insoluble from soluble sequences, as well as establish upper and lower bounds for LCST behavior at physiologically relevant temperatures. We also notice some discriminating power for the presence or absence of LCST behavior using average residue solvation energies, which is influenced by the presence of charge and other amino acid interactions (*115*). In this space, there is apparent discrimination of classes of sequences among newly produced and previously published sequences (Figure 38 and Table 16). Among these sequences there is separation of soluble versus insoluble sequences, and sequences with LCST above and below physiological ranges. Polycations and zwitterions do not appear to follow the same patterns as uncharged sequences, as the additional complexities introduced by charge remains to be characterized.

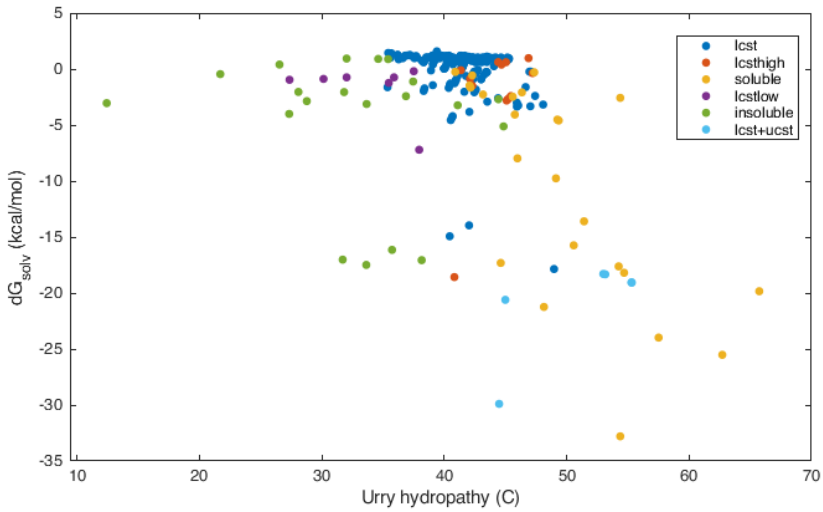

Figure 38: (A) Summary of ~60 newly produced sequences and of ~190 previously published repetitive proline and glycine-rich sequences, labeled as LCST, LCST with NaCl concentration (lcsthigh), LCST with urea (lcstlow), insoluble, soluble, and combined LCST and UCST (lcst+ucst). LCST sequences show clustering based on average residue solvation energies and Urry hydropathy.

We attempt to apply the general model of Chapter 2 for the prediction of T_t for sequences whose T_t measurements were unused during the development of the model, due to lack of variation in concentration and length and/or high nonaliphatic content, and for newly produced sequences of Chapter 4. Due to limited data for predicting θ_0 and θ beyond mostly-aliphatic sequences, we turn to existing amino-acid scales.

The Urry hydropathy scale delineates amino acids based on phase transition temperatures of ELPs of the form poly[VPGXG] where X is any guest residue (*14*). Due to the observation that many ELPs of this form have transition temperatures outside of physiological ranges, Urry produced variants of the form poly[$(VPGVG)_v(VPGXG)_x$], where v and x are mole fractions $v+x = 1$. Under the assumption of linearity of T_t upon substitution of amino acids, he used linear extrapolations for each amino acid in the T_t versus x space to obtain T_t and $x = 1$. The linearity assumption may hold for aliphatic

amino acids and temperature deviations within physiological ranges. For example, the lysine fractions for ELP data in this set do not surpass 0.15. At a periodicity of one per ~30 residues, charge-charge repulsion is insignificant (*14*). While Urry did not measure his transitions at θ point, his polymers had high molecular weights (> 50 kDa) and concentrations (40 mg/mL), and are therefore close to θ. Indeed, we found good correlation between our linear coefficients for θ_{linear} and the Urry hydropathy scale (R^2 = 0.952) for our set of mostly-aliphatic amino acids excluding Proline.

Therefore, we decided to use Urry hydropathy as a tool to predict θ for a more general range of sequences, assuming the majority of amino acids are aliphatic. As $\theta/(\theta-\theta_0)$ varies linearly with sequence, we used a linear fit of $\theta/(\theta-\theta_0)$ versus $H/(H-272.63)$, where H is the average Urry hydropathy in kelvin. Fitting gives the following linear equation (R^2 = 0.940),

$$\frac{\theta}{\theta-\theta_0} \approx 8.648\left(\frac{\sum_{AA} T_{\text{Urry},AA} f_{AA}}{\sum_{AA} T_{\text{Urry},AA} f_{AA} - 272.63}\right) - 53.72\,, \qquad \textbf{(49)}$$

where the sum term is the average Urry hydropathy for a sequence, the total sum of Urry hydropathies for each amino acid multiplied the fraction of amino acids in the sequence (Figure 39).

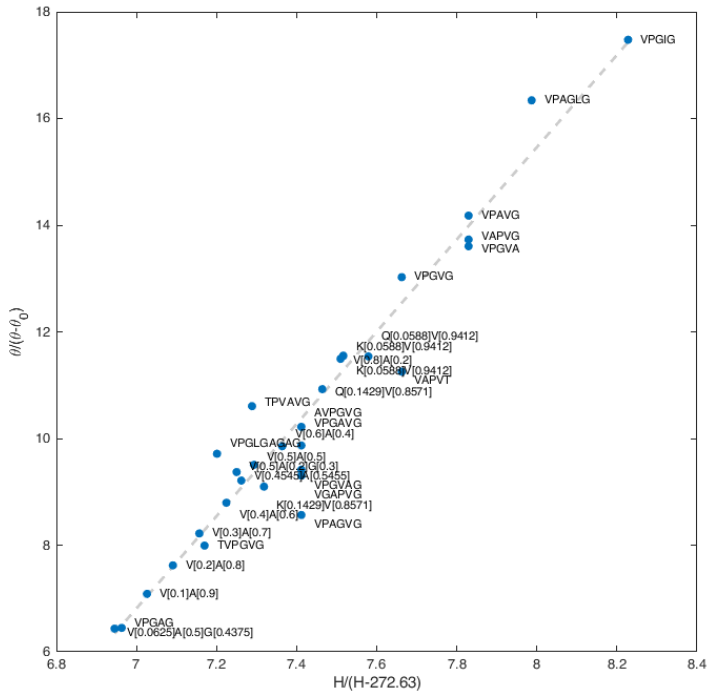

Figure 39: Linear prediction of fitted θ using Urry hydropathy. Dotted line indicates linear regression. Labels indicate the sequence pattern for each independently fit data set.

Average Urry hydropathy was used to predict $\theta/(\theta-\theta_0)$ for our new data set using a value of θ_0 = 272.63 K. The predictions were poorly correlated with measured values (R^2 = 0.22), likely due to variations in θ_0 and other variables due to differences in hydration and chain-chain interaction, as well as wide deviations in sequence diversity. As we make major deviations in sequence from canonical ELP, further understanding of model parameters such as θ_0 and θ are necessary to generalize the model predictions to a wider range of nonaliphatic sequences with diverse compositions and orderings.

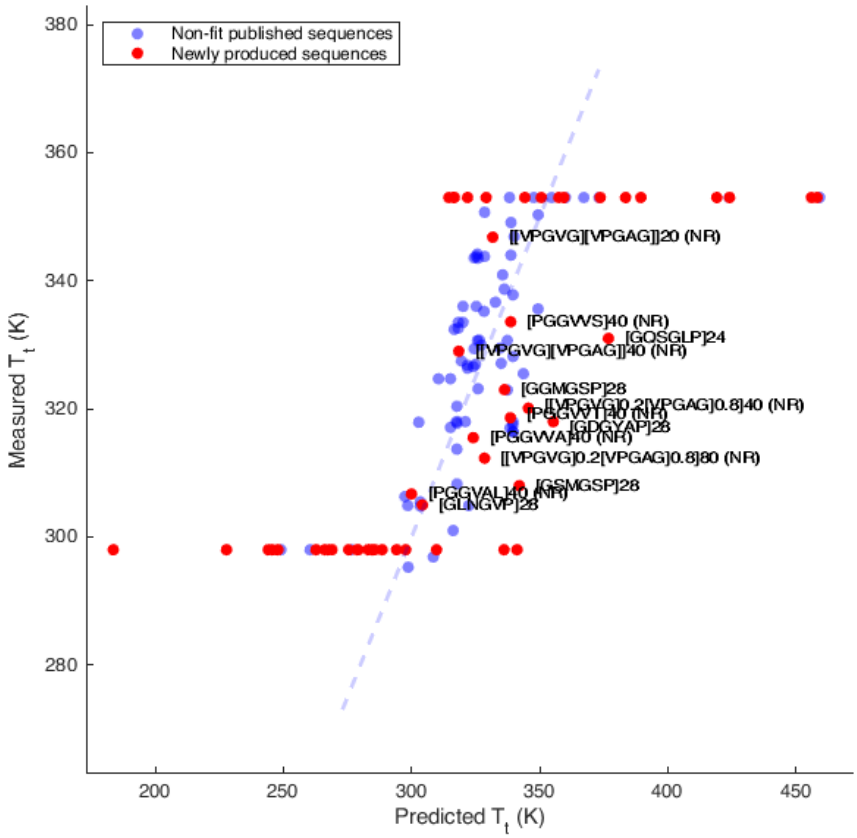

Figure 40: Comparison of measured T_t and predicted T_t using the LCST IDPP phase transition model with estimated values of θ and θ_0 = 272.63. Dense data points at measured T = 298 K and T = 353 K indicate insoluble and soluble IDPPs respectively that did not phase transition within physiologically relevant ranges. NR indicates nonrepetitive sequences.

However, we have found that Urry hydropathy and predictions of θ can be useful for discriminating soluble and insoluble IDPP sequences, as can be seen rank ordering of sequences by their Urry hydropathies (Table 9). Classification of this class of IDPPs can be achieved using a threshold Urry hydropathy of approximately 311 K.

Table 9: Purified protein polymer sequences. Proteins are labeled as insoluble, soluble, LCST, or LCST in the presence of Urea (lcstlow).

Sequence	Urry hydropathy (K)	Solubility/LCST
[GYGYAP]24	285.43558	insoluble

[GYGHAP]28	292.16666	insoluble
[PGGVVY]40 (NR)	294.70968	insoluble
[PGGVVF]40 (NR)	299.54839	insoluble
[GYSGVP]28	301.09444	insoluble
[GYVGTP]24	301.77914	insoluble
[GHIGVP]24	302.33333	insoluble
[GHHGTP]28	304.6667	insoluble
[PGGIKY]40 (NR)	304.70968	insoluble
[GGYSAP]28	304.82778	insoluble
[PGGVIL]40 (NR)	305.03226	insoluble
[PGGVFS]40 (NR)	305.03226	lcstlow
[GYMGKP]24	306.63804	insoluble
[GGLHTP]28	307.1667	insoluble
[PGGVVL]40 (NR)	307.6129	insoluble
[PGGVVI]40 (NR)	308.41935	insoluble
[GGMLAPGGMLAPGGMLAPGGMFAP]6	308.47239	lcstlow
[PGGLKF]40 (NR)	308.74194	insoluble
[PGGAFT]40 (NR)	308.90323	lcstlow
[GGMLAP]24	310.46626	insoluble
[PGGILT]40 (NR)	310.51613	lcstlow
[GGSHVP]28	310.6667	insoluble
[PGGVALKTIY]20 (NR)	310.95968	lcstlow
[PGGAKY]40 (NR)	311.16129	insoluble
[PGGVAL]40 (NR)	311.48387	lcst
[GLNGVP]28	312.07182	lcst
[PGGVIS]40 (NR)	313.90323	soluble
[[VPGVG]0.2[VPGAG]0.8]40 (NR)	314.1699	lcst
[[VPGVG]0.2[VPGAG]0.8]80 (NR)	314.26355	lcst
[PGGVVA]40 (NR)	314.87097	lcst
[GANMPPGASIPPGANIPPGASIPP]8	315.07109	soluble
[GANMPPGASIPPGANIPPGASIPP]7	315.14973	soluble
[GAGAIPGAGAIPGAGAIPVAGAIP]6	315.28834	soluble
[[VPGVG][VPGAG]]20 (NR)	315.56796	lcst
[[VPGVG][VPGAG]]40 (NR)	315.68227	lcst
[GANMPQGASIPPGANIPPGASIPP]6	316.17178	soluble
[GTHGTP]24	316.3333	soluble
[PGGVVS]40 (NR)	316.48387	lcst
[PGGVVT]40 (NR)	316.48387	lcst
[GKFGTP]24	317.61963	soluble
[GLSGSP]28	318.59893	soluble
[GNGNVP]24	318.76829	soluble
[PGGVALKTI]22.2	318.98387	soluble
[GIGQAP]28	319.34759	soluble
[GGMGSP]28	319.65193	lcst
[GQSGLP]24	320.03681	lcst

[PGGVAT]40 (NR)	320.35484	soluble
[GSMGSP]28	320.42541	lcst
[GIRGNPRLRGTPEYPGTPGIRGTP]7	321.14439	soluble
[GDGYAP]28	321.97238	lcst
[GGKHAP]24	322	soluble
[GGKHAP]28	322	soluble
[PGGVALKT]25 (NR)	322.12903	soluble
[GAGAIPGAEAIPGAGAIPGAGAIP]6	322.35583	soluble
[PGGVVK]40 (NR)	323.58065	soluble
[GRRGSP]28	327.38333	soluble
[GAQGSP]28	327.38333	soluble
[GGKHAPGGKHAPGGKHAPDDKHAP]6	331.5833	soluble

7. Conclusions

With the models and technologies developed in this work, we will continue to iterate the sequence space expansion pipeline to provide a feed of sequence data for our model-based pipeline. Discovery, combined with physics-based prediction, comprise a systematic exploration of the amino acid sequence space for the complete diversity of IDPPs with LCST behavior. As a technology, the pooled library approach to IDPP discovery will accelerate the global search of the amino acid sequence space. New design rules that emerge from this study can be used to design of novel and diverse sequences that can be easily synthesized using our previously published codon scrambling technology (*40*).

To reinforce the model and verify its validity and predictive power, measured values for the temperature dependence of R_g would be advantageous. Small angle X-ray scattering (SAXS) is a powerful method for characterization of intrinsically disordered proteins in solution, from which R_g can be calculated (*116*). Several studies have used SAXS to monitor changes in R_g with changes in temperature (*117*, *118*).

A furthered understanding of sequence effects on phase transition and surface tension requires rigorous characterization. Tedious measurements of transition temperatures of high length and high concentration libraries are necessary for each construct, which would allow for reliable extrapolations of model parameters such as θ_0 and θ. This may be feasible with the SYPRO orange assay because of its high throughput low volume requirements. Since canonical ELP is known to be well-hydrated, the most straightforward experiments on phase transition would be to repeat Urry's experiments on T_t and heat on sets of ELPs, except with a focus on concentration and length gradients in order to estimate θ_0 and $\theta/(\theta-\theta_0)$. With codon scrambling and major advances in high throughput gene synthesis, it would be possible to generate long and precise lengths of gene sequences for this data set as starting materials.

It is very possible that Urry's assumption of linearity for his application of extrapolations of amino acid changes would not be sufficient for major deviations from canonical ELP. Therefore, in order avoid large extrapolations in the sequence space, and to bring transition temperatures of highly soluble or insoluble ELP variants within range, salts such as Urea and NaCl would be needed to elevate or depress transition temperatures, respectively. While this adds additional dimensionality and complexity to the system, as the effects of salt on transition temperature can be sequence-context dependent and nonlinear, we believe that this additional dimensionality can be fully characterized, given the 96-well throughput of the SYPRO Orange assay.

Alternatively, pooled library synthesis via RCA opens the doors to interesting 'focused' experiments on local manifolds of the global sequence space. For example,

degenerate oligonucleotides with wobble bases could be used to vary the codons of only one of the amino acids in each motif; for example, in poly[GXGVP] where X is an amino acid coded by a degenerate codon. The degenerate bases can be chosen based on codons that encode amino acid sets with specific properties such as charge (RRR) or aliphalicity (DYD). The resulting proteins, which vary in distribution and/or fractional content of amino acids throughout the length of the sequence, could then be characterized in a high throughput format, allowing for specific trajectories in the sequence space.

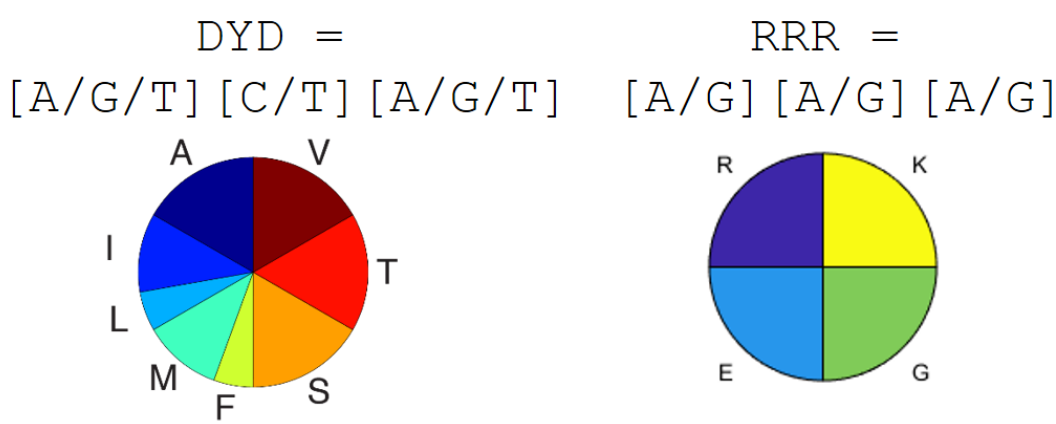

Figure 41: Distribution of amino acids encoded by degenerate codons, DYD and RRR. DYD represents three wobble bases which encode mostly aliphatic amino acids. RRR encode mostly charged amino acids.

A thermodynamic interpretation of the parameters of the model has yet to be made. Urry noted that virtually every change in the composition of the polypeptide and in the composition of the aqueous medium changes both the temperature and heat of phase transition, using the relation at transition,

$$\Delta H\left(\text{reference}\right)-\Delta H\left(\chi\right)=-\delta\Delta H\left(\chi\right)\approx\Delta G\left(\chi\right)+T\left(\text{G}\chi\text{GVP}\right)\delta\Delta S\left(\chi\right), \quad \textbf{(50)}$$

where χ refers to a change of any experimental variable including addition of salt or the substitution of a reference amino acid residue (Gly) by any other amino acid residue (*14*). An analogous thermodynamic treatment by Heyda and Dzubiella (*119*), has been used to describe the influence of Hofmeister salts on ELP phase transition temperature (*120*). This model relates phase transition temperature to concentration of cosolutes such as Guanidinium by using a Taylor series expansion of the free energy per monomer of coil to globule collapse with respect to perturbations in concentration of cosolutes and temperature. Solving for temperature change,

$$\Delta T = -\frac{mc_s + \frac{1}{2}m'c_s^{\,2}}{\Delta S_0 + \Delta S_0' c_s}, \qquad \textbf{(51)}$$

where c_s is the concentration of the cosolute, $\Delta S_0 = -(\partial \Delta G/\partial T)c_s$, $m = -(\partial \Delta G/\partial c_s)_T$., $m' = -(\partial^2 \Delta G/\partial c^2)_{T0,0}$, and $\Delta S_0' = -(\partial^2 \Delta G/\partial c \partial T)_{T0,0}$. This quadratic effect on LCST transition temperature has also been observed and modeled empirically (*121*) with salts like Guanidinium, which is attributed here to nonzero m'.

A potential future direction would be to analogously relate the substitution of amino acids on a peptide backbone to thermodynamic parameters. Like salt, amino acid changes affect the solvent quality at phase transition. A similar treatment to that of Heyda and Dzubiella would be to relate the following thermodynamic properties to amino acids to transition temperature: free energy change upon introduction of a sidechain to the IDPP backbone, entropy of backbone, change in first derivative of free energy upon introduction of sidechain, and entropy change upon introduction of sidechain. The thermodynamic quantities for phase transition upon amino acid

substitution of IDPPs are currently unknown. However, it is worth noting that Urry found a near linear correlation between T_t and enthalpy of transition for aliphatic amino acids suggesting that transition entropy differences among them are minimal (*14*).

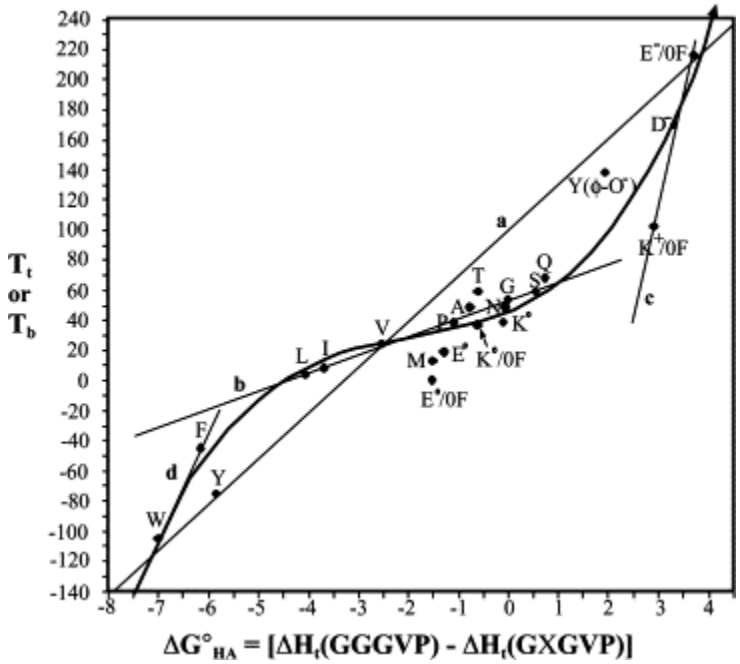

Figure 42: Urry's relationship between transition T_t and change transition heat upon amino acid substitution near θ. Linear and cubic fits indicate families of amino acids which may have shared change in entropy upon amino acid substitution. From (*14*).

Phase transitions and criticality appear in many physical systems, ranging in scale from ferromagnetic systems to protein folding to cosmological phase transitions after the big bang (*122*, *123*). It is believed that near phase transition, the thermodynamic properties of these systems depend on shared scale-invariant features such as dimensionality and symmetry. The fact that proteins such as elastin are poised close to their critical point suggests that over time, natural selection has had the chance to exploit these features.

Our proposed model framework also poses interesting questions on the entropic and energetic nature of θ and θ_0 whose details remain to be understood, with implications towards understanding surface tension and hydrophobicity of protein globules (*124*, *125*). The importance of understanding the physics of IDP interactions with water should not be discounted. IDPs are considered the "dark matter" of biology, yet they can be considered one of the most fundamental objects (*126*). From a reductionist approach, it may hold the key to understanding the energetic and entropic contributions of amino acids in the context of a protein. To reiterate a point made by Urry, the complexity of biology wanes as biologically relevant variables couple through their thermodynamic parameters (*14*). The challenge of achieving complete understanding of the energetics of a complex protein was once thought to be nigh impossible. As the field of polymer physics matures, its interface with biology in the next decade may lead us to the next major advance towards undertaking this challenge (*127*).

Appendix

Table 10: Motifs of 1020 microarray-synthesized oligonucleotides

PGAAGA	PGDDGA	PGWGGA	PGWGFA	PGDGDR	PGWGGR	PGWFGR	PGEGNI	PGGNMV	PGGDGI	PGDGWT	PGGHSQ	PGQWGT	PGGTEI	PGGGGK
PGAGRA	PGQGDA	PGGGYA	PGGFYA	PGGRDQ	PGYGGR	PGYGFR	PGNEGL	PGGNFS	PGGDGL	PGDGYT	PGGQHT	PGYQGT	PGEGWI	PGGGGS
PGGNAA	PGDGEA	PGGAGV	PGAFGV	PGEDGR	PGGGVR	PGFRGV	PGGENK	PGGFNT	PGGDGK	PGDTGV	PGGHWQ	PGGTVQ	PGGEYI	PGGGGT
PGDAGA	PGGDGA	PGGHHA	PGSGAS	PGGGDR	PGGHHR	PGSRGS	PGGNME	PGFNGV	PGGDGS	PGGWDV	PGGYHQ	PGGWWQ	PGVEGI	PGGGGV
PGGAQA	PGGHDA	PGAHGI	PGGTAS	PGHGDR	PGRGHI	PGGSTR	PGGNFE	PGNGSS	PGGDGT	PGDYGV	PGHQGV	PGYWGQ	PGELGL	PGGHGI
PGEAGA	PGADGI	PGHGAL	PGGWSA	PGRGDI	PGGHRL	PGGRWS	PGGNES	PGNSGT	PGGGDV	PGDVGV	PGIGQI	PGGVWQ	PGLKGE	PGGHGL
PGGAGA	PGADGL	PGGKHA	PGGYSA	PGDLGR	PGHRGK	PGYGRS	PGGNET	PGGNWS	PGHDGI	PGQQGQ	PGGLQI	PGYYGQ	PGMLGE	PGHGGK
PGAGHA	PGKGDA	PGGHMA	PGGSAV	PGDKGR	PGHMGR	PGRVGS	PGGWNE	PGYGNS	PGGDHL	PGGQQE	PGGQKI	PGGQYV	PGGFEL	PGGHGS
PGAGAI	PGGMDA	PGFGHA	PGGTAT	PGGMDR	PGFGHR	PGRGTT	PGYNGE	PGVGNS	PGHGDK	PGQGGQ	PGGQMI	PGVQGV	PGELGS	PGHGGT
PGAGAL	PGFGDA	PGHGAS	PGGWAT	PGGFDR	PGHGRS	PGRWGT	PGEGNV	PGNTGT	PGHGDS	PGGHQQ	PGIFGQ	PGGEEE	PGLEGT	PGHGGV
PGGKAA	PGGADS	PGHGAT	PGYAGT	PGRDGS	PGTHGR	PGRYGT	PGGGNI	PGGWNT	PGHDGT	PGQIGQ	PGGSQI	PGGGEE	PGWEGL	PGIGGI
PGAGMA	PGADGT	PGWHGA	PGAVGT	PGRDGT	PGWGHR	PGTVGR	PGGGNL	PGYGNT	PGGDHV	PGGLQQ	PGQGTI	PGEHGE	PGYLGE	PGGLGI
PGAGFA	PGGDWA	PGYGHA	PGGWWA	PGDWGR	PGYHGR	PGWGWR	PGGNGK	PGNVGT	PGGDII	PGGQKQ	PGGQWI	PGEGEI	PGLGVE	PGKGGI
PGAGAS	PGDGYA	PGGHVA	PGYGWA	PGGYDR	PGHRGV	PGWYGR	PGGNGS	PGWGNV	PGGLDI	PGQMGQ	PGQGYI	PGELGE	PGEKGK	PGGMGI
PGGTAA	PGGDAV	PGAIGI	PGGWVA	PGRGDV	PGIGIR	PGGWVR	PGGNGT	PGYNGV	PGGDKI	PGFQGQ	PGQGIV	PGEKGE	PGGKME	PGGFGI
PGWAGA	PGGQQA	PGGLAI	PGYGYA	PGQGQR	PGLRGI	PGYYGR	PGGGNV	PGGNVV	PGDGMI	PGQGSQ	PGQLGL	PGGMEE	PGGFEK	PGGIGS
PGGAYA	PGEGAQ	PGGKAI	PGVGYA	PGQRGE	PGGKRI	PGGRYV	PGGHNI	PGDDGQ	PGGDFI	PGTQGQ	PGKQGL	PGFEGE	PGGKES	PGGGTI
PGVAGA	PGGGQA	PGAMGI	PGVGVA	PGRGGQ	PGGIMR	PGGRVV	PGNHGL	PGDGDE	PGGIDS	PGWGQQ	PGQGML	PGESGE	PGGKET	PGGGWI
PGGRAR	PGGHQA	PGFGIA	PGRRGR	PGGHRQ	PGFGRI	PGNNGQ	PGGHNK	PGDDGI	PGGTDI	PGYGQQ	PGFLGQ	PGEEGT	PGKWGE	PGGGYI
PGANGR	PGIGQA	PGSGAI	PGGRNR	PGIQGR	PGRIGS	PGGNNE	PGGNHS	PGDGDL	PGGDWI	PGVQGQ	PGQSGL	PGEWGE	PGYKGE	PGGIGV
PGGADR	PGAQGL	PGGTIA	PGGRDR	PGGLQR	PGIRGT	PGGNNI	PGHNGT	PGGDDK	PGGDYI	PGQEGE	PGTLGQ	PGYEGE	PGEGKV	PGLGGL
PGRGQA	PGGKQA	PGAGWI	PGQGRR	PGQRGK	PGGWRI	PGGNNL	PGGNHV	PGDDGS	PGGIDV	PGGEGQ	PGGLWQ	PGGEEV	PGGMME	PGGKGL
PGAEGR	PGGQMA	PGAYGI	PGRGRE	PGMQGR	PGYGRI	PGNNGK	PGNIGI	PGDDGT	PGDGLL	PGHEGQ	PGQYGL	PGGGGE	PGGMFE	PGMGGL
PGGRGA	PGFGQA	PGAIGV	PGGGRR	PGQGFR	PGIGVR	PGNNGS	PGGLNI	PGGDDV	PGDGKL	PGGEQI	PGQLGV	PGGGHE	PGMEGS	PGGGFL
PGRGHA	PGAQGS	PGLGLA	PGGRHR	PGRQGS	PGLRGL	PGGNNT	PGGKNI	PGDQGQ	PGMDGL	PGQGLE	PGKKGQ	PGIGGE	PGTGME	PGGLGS
PGIGAR	PGAGQT	PGKGLA	PGRRGI	PGRGQT	PGRKGL	PGNGNV	PGGNMI	PGGQDE	PGFGDL	PGGKQE	PGMQGK	PGGEGL	PGGWME	PGGGLT
PGRGLA	PGWGQA	PGGMLA	PGGLRR	PGQWGR	PGRGML	PGDNGQ	PGGNFI	PGGDGQ	PGDLGS	PGGMEQ	PGGFQK	PGGEGK	PGMGYE	PGGGWL
PGGRKA	PGQGYA	PGFGAL	PGKRGR	PGYGRQ	PGFRGL	PGGDNE	PGSNGI	PGGHDQ	PGGTDL	PGQFGE	PGQGSK	PGGMGE	PGGVME	PGGGYL
PGGMRA	PGQGAV	PGGSAL	PGRGMR	PGQVGR	PGGRSL	PGDGNI	PGNIGT	PGGIDQ	PGWDGL	PGQEGS	PGGQKT	PGGGFE	PGFFGE	PGGVGL
PGFGRA	PGGAEE	PGGATL	PGFRGR	PGERGE	PGRGLT	PGGNDL	PGGNWI	PGQGDL	PGDGYL	PGTEGQ	PGGKWQ	PGGGSE	PGEFGS	PGGKGK
PGAGSR	PGGEGA	PGWGLA	PGRRGS	PGEGGR	PGRWGL	PGNDGK	PGNYGI	PGKDGQ	PGLGDV	PGGEWQ	PGKYGQ	PGEGGT	PGGFET	PGGGMK
PGRGAT	PGHGEA	PGYGLA	PGTRGR	PGHEGR	PGGLYR	PGDNGS	PGGNIV	PGGDMQ	PGKGDK	PGGQYE	PGQGKV	PGWGGE	PGWGFE	PGGFGK
PGWGAR	PGAEGI	PGALGV	PGRGWR	PGIEGR	PGGRVL	PGGNDT	PGGNLL	PGFDGQ	PGGMDK	PGQGEV	PGGMMQ	PGYGGE	PGYFGE	PGSGGK
PGRGYA	PGAGEL	PGGAKK	PGGRYR	PGERGL	PGKRGK	PGDNGV	PGGKNL	PGSDGQ	PGFDGK	PGGGGQ	PGGFMQ	PGGEGV	PGGEFV	PGGKGT
PGVGRA	PGEGKA	PGGKMA	PGRRGV	PGEKGR	PGGKMR	PGQGNQ	PGGNML	PGDQGT	PGGKDS	PGHGGQ	PGMQGS	PGGHHE	PGSEGS	PGGWGK
PGGNNA	PGAGME	PGGAFK	PGGNNR	PGGEMR	PGGKFR	PGNEGQ	PGFNGL	PGDWGQ	PGKDGT	PGGIGQ	PGQMGT	PGEHGI	PGGSET	PGGGYK
PGDNGA	PGGFEA	PGGSKA	PGGDNR	PGRGFE	PGGSKR	PGNGGQ	PGGNSL	PGDYGQ	PGGWDK	PGGQGL	PGGMWQ	PGHEGL	PGEGWS	PGGKGV
PGQGNA	PGGEAS	PGKGTA	PGNRGQ	PGGRES	PGRKGT	PGHNGQ	PGLNGT	PGVDGQ	PGGDYK	PGGKGQ	PGYMGQ	PGGEHK	PGYEGS	PGGMGS
PGAGNE	PGGATE	PGGAWK	PGNRGE	PGGRTE	PGGWKR	PGQNGI	PGWNGL	PGGEDE	PGDGKV	PGGGMQ	PGVMGQ	PGHGME	PGVEGS	PGGMGT
PGGGNA	PGGWAE	PGAGYK	PGGGNR	PGEGWR	PGKGYR	PGNQGL	PGGNYL	PGDGGE	PGDMGS	PGFGGQ	PGGFFQ	PGHFGE	PGGTET	PGMGGV
PGNGHA	PGGAYE	PGVGAK	PGHGNR	PGGEYR	PGKGVR	PGNGKQ	PGLNGV	PGDHGE	PGDMGT	PGSGGQ	PGSFGQ	PGHEGS	PGWEGT	PGGFGS
PGANGI	PGAGEV	PGGMMA	PGNRGI	PGEVGR	PGGMMR	PGGNMQ	PGKNGK	PGEDGI	PGGMDV	PGGTGQ	PGGQFT	PGHEGT	PGGYTE	PGGFGT
PGANGL	PGGGGA	PGGMFA	PGGRNL	PGGGGR	PGGMFR	PGNFGQ	PGGNMK	PGLDGE	PGGFDS	PGWGGQ	PGGFWQ	PGWHGE	PGGETV	PGFGGV
PGAGNK	PGGGHA	PGGMAS	PGNKGR	PGGGHR	PGGSMR	PGGNQS	PGNFGK	PGEGDK	PGFDGT	PGYGGQ	PGFYGQ	PGHGYE	PGWGWE	PGGGSS
PGGNMA	PGGGAI	PGAMGT	PGGMNR	PGGGIR	PGRMGT	PGNQGT	PGGNKS	PGDGME	PGGFDV	PGGGQV	PGFQGV	PGGEHV	PGYWGE	PGSGGT
PGFGNA	PGGGLA	PGGMWA	PGFGNR	PGGGLR	PGWGMR	PGGNWQ	PGGNKT	PGFDGE	PGDSGS	PGHHGQ	PGSQGS	PGIGEI	PGWEGV	PGGWGS
PGGNSA	PGGGAK	PGMYGA	PGRNGS	PGGGKR	PGMYGR	PGGNYQ	PGGWNK	PGDEGS	PGSDGT	PGGQHI	PGGQST	PGELGI	PGGYYE	PGYGGS
PGGTNA	PGGGMA	PGMGVA	PGNRGT	PGGGMR	PGMRGV	PGVNGQ	PGNYGK	PGGDTE	PGDGWS	PGHQGL	PGQWGS	PGGKEI	PGGEYV	PGSGGV
PGWGNA	PGFGGA	PGFGFA	PGGNWR	PGFGGR	PGFGFR	PGGENE	PGGNKV	PGGWDE	PGGYDS	PGHQGK	PGYQGS	PGEGMI	PGEVGV	PGTGGT
PGGNYA	PGAGGS	PGAFGS	PGNGYR	PGGRGS	PGGFSR	PGGNGE	PGNMGS	PGGDYE	PGDVGS	PGHMGQ	PGQGSV	PGEGFI	PGGGGI	PGGGWT
PGANGV	PGGAGT	PGFGTA	PGRNGV	PGGGTR	PGRFGT	PGGNHE	PGGMNT	PGGDVE	PGTGDT	PGFHGQ	PGQGTT	PGEGSI	PGGGGL	PGGYGT
PGTGGV	PGHTGV	PGYIGT	PGYLGV	PGMYGT	PGWHGI	PGGLFI	PGGKKL	PGFFGK	PGWFGV	PGGHWK	PGGTMI	PGGTFL	PGGVWK	PGGVVS
PGWGGV	PGGWHV	PGVGTI	PGGVLV	PGMVGT	PGGYHI	PGGLSI	PGGMKL	PGGKFS	PGFYGV	PGYHGK	PGWGMI	PGWFGL	PGYYGK	PGTTGT
PGGGYV	PGYHGV	PGGWWI	PGKKGK	PGWMGV	PGHIGV	PGGLTI	PGFGKL	PGKFGT	PGFVGV	PGGKHV	PGMYGI	PGYFGL	PGYVGK	PGTWGT
PGGGVV	PGGVHV	PGGWYI	PGMKGK	PGGMYV	PGGLHL	PGIGWL	PGKSGL	PGFWGK	PGSGSS	PGHMGS	PGIGMV	PGFGVL	PGGKVV	PGYTGT
PGHHGI	PGIIGI	PGWIGV	PGKFGK	PGGVMV	PGKGHL	PGGLYI	PGGKTL	PGYFGK	PGSTGS	PGHMGT	PGFGFI	PGLSGS	PGMMGS	PGVTGT
PGHHGL	PGGLII	PGYYGI	PGKKGS	PGFFGS	PGHMGL	PGVIGL	PGGWKL	PGGKFV	PGSGWS	PGHMGV	PGSFGI	PGSLGT	PGMMGT	PGGWWT
PGGHHK	PGGKII	PGGIYV	PGTKGK	PGFFGT	PGHFGL	PGIKGK	PGKGYL	PGSKGS	PGGSYS	PGHGFS	PGTFGI	PGWLGS	PGMMGV	PGWYGT
PGHGHS	PGGIMI	PGVGIV	PGGWKK	PGGFFV	PGHSGL	PGGKMI	PGKGVL	PGGTKS	PGSSGV	PGFHGT	PGFWGI	PGYLGS	PGMFGS	PGWVGT
PGHHGT	PGIFGI	PGLLGL	PGKYGK	PGFGSS	PGGLHT	PGFGKI	PGGMML	PGKWGS	PGTSGT	PGHFGV	PGFYGI	PGSVGL	PGMFGT	PGYYGT
PGHHGV	PGIGSI	PGLKGL	PGGKKV	PGSGFT	PGGHWL	PGGSKI	PGFMGL	PGYGKS	PGGTWS	PGSHGS	PGGIFV	PGGTTL	PGFGMV	PGYVGT
PGHGII	PGGIIT	PGGMLL	PGGMMK	PGGFWS	PGGHYL	PGGKTI	PGMLGS	PGKVGS	PGGSYT	PGSHGT	PGSIGS	PGGWTL	PGSMGS	PGGTVV
PGGHIL	PGWIGI	PGFLGL	PGMFGK	PGYFGS	PGGLHV	PGWGKI	PGGTML	PGGKTT	PGVGTS	PGHGWS	PGGITS	PGTYGL	PGSMGT	PGWWGV
PGGKHI	PGGIYI	PGLGLS	PGGMSK	PGFVGS	PGKHGK	PGGKYI	PGGWML	PGGTWK	PGWGWS	PGHYGS	PGGSWI	PGLVGT	PGGMWS	PGGWYV
PGGHMI	PGVIGI	PGTLGL	PGMKGT	PGTFGT	PGHGMK	PGGKVI	PGMYGL	PGYKGT	PGWYGS	PGGSHV	PGYIGS	PGWWGL	PGMYGS	PGVGWV
PGHFGI	PGILGL	PGGLWL	PGGMWK	PGWFGT	PGGHFK	PGMGMI	PGGMLV	PGGVTK	PGWVGS	PGTHGT	PGGISV	PGWYGL	PGGMSV	PGGYYV
PGGHSI	PGIGKL	PGLYGL	PGYMGK	PGYFGT	PGGKHS	PGFGMI	PGFFGL	PGWWGK	PGGYYS	PGHWGT	PGTIGT	PGVGWL	PGTMGT	PGGYVV
PGHGTI	PGGIML	PGLLGV	PGKGMV	PGTFGV	PGGKHT	PGMIGS	PGLGFS	PGGWYK	PGYSGV	PGGYHT	PGWGTI	PGYGYL	PGWMGT	PGVVGV

Table 11: Designed nucleotide sequences of 1020 microarray-synthesized oligonucleotides

1. GGTGCAGCGGGAGCGCCAGGAGCTGCTGGAGCCCCAGGGGCAGCCGGTGCTCCGGGTGCCGCAGGCGCACCA	511. GGTGATGGGTGGACACCGGGAGATGGTTGGACCCCAGGAGACGGATGGACTCCAGGCGATGGCTGGACACCA
2. GGTGCGGGACGAGCGCCAGGAGCCGGTAGAGCACCCGGAGCAGGTCGTGCCCCTGGAGCTGGGAGAGCACCA	512. GGTGATGGTTATACCCCAGGAGATGGATATACTCCGGGAGACGGTTACACTCCAGGCGATGGCTATACACCA
3. GGTGGCAATGCAGCCCCAGGGGGTAACGCCGCTCCTGGAGGAAATGCTGCTCCGGGAGGTAATGCGGCACCA	513. GGTGATACTGGTGTGCCCGGAGACACTGGGGTTCCTGGCGATACGGGAGTCCCTGGAGATACAGGCGTACCA
4. GGTGATGCGGGTGCTCCAGGAGATGCCGGAGCCCCAGGGGATGCAGGCGCTCCCGGTGACGCTGGCGCACCA	514. GGTGGTTGGGATGTGCCAGGAGGGTGGGACGTACCCGGCGGATGGGATGTTCCAGGGGGCTGGGATGTACCA
5. GGTGGGGCTCAAGCACCCGGAGGCGCTCAGGCTCCTGGCGGTGCCCAAGCTCCAGGGGGAGCACAAGCACCA	515. GGTGATTACGGTGTTCCAGGAGATTATGGAGTGCCCGGAGACTATGGGGTCCCTGGCGACTACGGCGTACCA
6. GGTGAAGCGGGTGCGCCTGGGGAAGCCGGTGCTCCGGGAGAGGCAGGAGCCCCTGGAGAAGCTGGCGCACCA	516. GGTGATGTAGGGGTCCCAGGCGATGTCGGTGTGCCCGGAGACGTAGGAGTTCCTGGAGATGTTGGCGTACCA
7. GGTGGAGCGGGTGCGCCTGGAGGTGCCGGTGCTCCGGGAGGGGCAGGAGCCCCAGGGGGAGCTGGCGCACCA	517. GGTCAGCAGGGACAACCCGGTCAACAGGGGCAGCCAGGACAGCAAGGCCAGCCTGGGCAACAAGGTCAACCA
8. GGTGCAGGCCATGCGCCAGGAGCTGGACATGCTCCGGGAGCCGGTCATGCCCCTGGAGCGGGTCACGCACCA	518. GGTGGGCAACAAGAGCCTGGAGGTCAGCAAGAACCCGGCGGTCAACAGGAGCCAGGGGGACAGCAGGAACCA
9. GGTGCCGGTGCAATCCCAGGGGCTGGAGCTATTCCAGGAGCGGGAGCCATACCCGGAGCAGGCGCAATACCA	519. GGTCAAGGCGGCCAACCGGGTCAGGGAGGCAGCCAGGACAAGGTGGACAGCCTGGGCAAGGAGGTCAACCA
10. GGTGCAGGAGCTTTGCCTGGGGCTGGCGCACTTCCTGGAGCGGGAGCATTACCCGGAGCCGGTGCTTTACCA	520. GGTGGGCACCAACAGCCTGGAGGACATCAACAACCCGGCGGTCATCAGCAGCCAGGGGGACACCAGCAACCA
11. GGTGGTAAAGCTGCGCCAGGAGGAAAAGCGGCTCCGGGTGGGAAAGCAGCCCCTGGAGGTAAGGCAGCACCA	521. GGTCAAATAGGGCAGCCAGGACAAATTGGACAGCCTGGGCAAATCGGCCAACCCGGTCAGATAGGTCAACCA
12. GGTGCGGGAATGGCACCCGGAGCTGGGATGGCCCCTGGAGCAGGCATGGCTCCTGGCGCTGGTATGGCACCA	522. GGTGGACTTCAACAGCCAGGAGGTTTACAACAACCCGGCGGTCTTCAGCAGCCTGGGGGATTACAGCAACCA
13. GGTGCAGGATTTGCTCCCGGAGCGGGTTTTGCGCCTGGAGCTGGCTTTGCCCCAGGAGCCGGTTTCGCACCA	523. GGTGGGCAAAAACAACCCGGCGGTCAGAAACAGCCAGGGGGACAAAAGCAGCCTGGAGGACAGAAGCAACCA
14. GGTGCAGGAGCATCTCCGGGTGCCGGTGCTTCCCCAGGGGCTGGCGCAAGTCCTGGAGCGGGAGCTTCACCA	524. GGTCAAATGGGGCAGCCAGGACAGATGGGACAACCCGGCCAAATGGGTCAGCCTGGACAAATGGGCCAACCA
15. GGTGGAACTGCTGCCCCTGGAGGTACAGCCGCTCCCGGAGGAACCGCAGCTCCTGGGGGTACTGCAGCACCA	525. GGTTTTCAGGGTCAACCGGGATTCCAGGGGCAGCCAGGATTTCAAGGACAGCCTGGGTTCCAAGGTCAACCA
16. GGTTGGGCCGGTGCTCCAGGGTGGGCAGGCGCTCCTGGCTGGGCGGGGGCTCCGGGATGGGCTGGTGCACCA	526. GGTCAAGGATCGCAGCCTGGACAGGGTTCTCAACCGGGGCAAGGTAGTCAGCCAGGACAAGGGAGTCAACCA
17. GGTGGAGCGTATGCCCCTGGGGGAGCTTATGCGCCAGGAGGAGCCTATGCACCCGGCGGAGCATATGCACCA	527. GGTACACAAGGGCAGCCTGGAACGCAAGGTCAACCGGGGACTCAAGGACAGCCAGGAACTCAGGGTCAACCA
18. GGTGTAGCTGGAGCGCCAGGAGTAGCGGGAGCCCCAGGGGTTGCCGGTGCTCCGGGTGTCGCAGGCGCACCA	528. GGTTGGGGGCAACAGCCTGGATGGGGACAACAACCCGGCTGGGGTCAGCAGCCAGGGTGGGGCCAACAACCA
19. GGTGGAAGAGCGAGACCCGGAGGTAGAGCTAGGCCAGGCGGCAGAGCACGTCCTGGGGGAAGGGCTCGACCA	529. GGTTATGGTCAGCAGCCTGGATATGGACAACAACCGGGGTATGGGCAACAGCCAGGATACGGACAGCAACCA
20. GGTGCGAATGGAAGACCCGGAGCAAATGGGCGTCCAGGAGCTAATGGTAGGCCAGGGGCTAACGGTCGACCA	530. GGTGTTCAAGGGCAGCCAGGAGTCCAAGGACAACCCGGCGTACAAGGTCAGCCTGGGGTACAGGGACAACCA
21. GGTGGGGCAGATCGTCCTGGAGGAGCCGATAGACCGGGTGGTGCTGATAGGCCAGGGGGAGCGGATCGACCA	531. GGTCAGGAAGGTGAGCCTGGACAAGAGGGAGAACCCGGTCAAGAAGGCGAGCCAGGGCAGGAGGGGGAACCA
22. GGTAGAGGCCAAGCTCCTGGAAGAGGACAAGCGCCGGGTCGTGGTCAAGCCCCAGGGAGAGGTCAGGCACCA	532. GGTGGGGAAGGACAGCCTGGAGGAGAAGGGCAACCCGGCGGAGAGGGTCAGCCAGGAGGTGAAGGTCAACCA
23. GGTGCCGAAGGGCGTCCTGGAGCGGAGGGTAGGCCAGGGGCTGAAGGAAGACCGGGAGCAGAAGGTCGACCA	533. GGTCACGAGGGACAGCCTGGACATGAGGGTCAGCCAGGGCATGAAGGACAACCCGGACACGAAGGTCAACCA
24. GGTGGAAGAGGTGCTCCCGGAGGCAGAGGAGCCCCAGGGGGTCGTGGCGCTCCAGGAGGTAGAGGGGCACCA	534. GGTGGTGAACAAATTCCTGGAGGAGAGCAAATCCCTGGGGGTGAGCAGATACCGGGCGGAGAACAGATACCA
25. GGTAGAGGTCACGCTCCGGGTCGTGGTCATGCCCCTGGGAGAGGACATGCTCCTGGAAGAGGCCATGCACCA	535. GGTCAAGGTTTGGAGCCAGGACAGGGACTAGAACCGGGGCAAGGATTAGAGCCTGGACAAGGGTTAGAACCA
26. GGTATTGGGGCTCGTCCAGGAATCGGAGCAAGGCCAGGGATAGGAGCTAGACCCGGCATAGGTGCGCGACCA	536. GGTGGGAAGCAGGAGCCTGGAGGAAAACAAGAACCGGGCGGAAAGCAAGAGCCAGGGGGTAAACAGGAACCA
27. GGTAGAGGTCTCGCTCCTGGAAGAGGACTTGCTCCCGGTCGTGGATTAGCCCCAGGCAGAGGGTTAGCACCA	537. GGTGGTATGGAACAGCCAGGAGGGATGGAGCAGCCTGGGGGCATGGAACAACCCGGCGGAATGGAACAACCA
28. GGTGGAAGAAAGGCTCCTGGCGGTCGTAAAGCCCCTGGAGGACGCAAAGCTCCCGGAGGTAGAAAAGCACCA	538. GGTCAATTCGGCGAGCCAGGACAATTTGGTGAACCCGGACAGTTTGGAGAGCCTGGGCAGTTCGGTGAACCA
29. GGTGGGATGAGAGCGCCAGGAGGTATGCGAGCCCCTGGAGGAATGCGTGCTCCGGGAGGCATGAGGGCACCA	539. GGTCAGGAAGGATCTCCAGGACAGGAGGGTTCGCCTGGGCAAGAAGGTAGTCCCGGTCAAGAGGGGTCACCA
30. GGTTTTGGTCGTGCTCCGGGATTTGGACGAGCGCCTGGATTCGGTAGAGCCCCAGGGTTTGGCAGAGCACCA	540. GGTACAGAAGGACAGCCAGGAACAGAGGGTCAACCGGGGACTGAAGGGCAGCCTGGAACGGAAGGTCAACCA
31. GGTGCCGGATCTCGTCCTGGAGCGGGTTCTAGGCCAGGGGCTGGGTCAAGACCGGGAGCAGGTAGTCGACCA	541. GGTGGTGAATGGCAACCCGGAGGAGAGTGGCAACCTGGGGGCGAATGGCAGCCTGGAGGGGAATGGCAACCA
32. GGTAGAGGTGCGACCCCAGGCAGAGGAGCTACTCCCGGACGTGGAGCAACGCCAGGAAGAGGGGCTACACCA	542. GGTGGGCAATATGAGCCTGGAGGACAGTATGAACCCGGCGGTCAGTACGAGCCAGGGGGACAATACGAACCA
33. GGTTGGGGAGCAAGACCCGGCTGGGGCGCTCGTCCAGGGTGGGGAGCTAGGCCAGGATGGGGTGCCCGACCA	543. GGTCAAGGTGAAGTACCCGGACAAGGAGAGGTTCCAGGCCAAGGCGAAGTTCCTGGGCAAGGGGAAGTACCA
34. GGTCGTGGCTATGCCCCAGGAAGAGGTTACGCTCCCGGTAGAGGGTATGCGCCTGGACGAGGATATGCACCA	544. GGTGGGGGAGGGCAGCCAGGGGGTGGTGGACAACCCGGCGGAGGAGGTCAGCCTGGAGGCGGTGGCCAACCA
35. GGTGTTGGACGAGCCCCTGGAGTCGGTCGTGCTCCCGGAGTGGGTAGAGCGCCAGGAGTAGGAAGAGCACCA	545. GGTCATGGGGGTCAGCCTGGACACGGCGGACAACCGGGACATGGAGGACAGCCAGGGCATGGTGGTCAACCA
36. GGTGGAAACAATGCCCCTGGAGGTAACAACGCTCCCGGAGGAAATAATGCTCCTGGCGGTAATAACGCACCA	546. GGTGGAATTGGTCAGCCAGGAGGTATAGGACAACCCGGCGGAATAGGGCAGCCTGGGGGAATCGGTCAACCA
37. GGTGATAACGGAGCCCCTGGGGACAATGGCGCTCCTGGAGATAATGGAGCACCGGGTGACAACGGCGCACCA	547. GGTGGTCAGGGTCTTCCAGGAGGTCAAGGTTTACCGGGCGGCCAAGGATTGCCTGGGGGACAAGGGTTACCA
38. GGTCAAGGTAATGCCCCAGGCCAAGGAAATGCTCCGGGTCAGGGTAACGCTCCTGGACAAGGGAATGCACCA	548. GGTGGGAAAGGTCAACCCGGCGGTAAGGGGCAGCCTGGAGGTAAAGGACAGCCAGGAGGAAAAGGCCAACCA
39. GGTGCCGGAAATGAGCCAGGAGCAGGGAATGAACCGGGAGCTGGAAACGAGCCTGGGGCGGGTAATGAACCA	549. GGTGGTGGAATGCAACCGGGCGGAGGTATGCAGCCAGGGGGTGGGATGCAACCTGGAGGAGGCATGCAACCA
40. GGTGGTGGAAATGCTCCCGGAGGGGGTAACGCTCCTGGAGGAGGTAATGCCCCAGGCGGTGGCAATGCACCA	550. GGTTTTGGAGGGCAGCCTGGGTTTGGTGGTCAACCGGGATTCGGCGGACAGCCAGGATTTGGGGGTCAACCA
41. GGTAACGGTCACGCTCCTGGAAATGGACATGCTCCGGGTAATGGTCATGCGCCAGGGAATGGGCATGCACCA	551. GGTAGCGGAGGGCAGCCTGGAAGTGGAGGTCAACCGGGTTCTGGGGGACAGCCAGGATCAGGTGGTCAACCA
42. GGTGCGAATGGAATCCCTGGAGCAAATGGCATACCCGGAGCTAACGGTATTCCTGGGGCTAATGGTATACCA	552. GGTGGTACAGGACAACCCGGCGGAACTGGTCAGCCTGGAGGTACTGGGCAGCCAGGGGGAACAGGTCAACCA
43. GGTGCGAATGGATTGCCAGGAGCTAACGGACTTCCAGGGGCTAATGGTCTACCGGGAGCAAATGGGTTACCA	553. GGTTGGGGAGGACAGCCAGGGTGGGGTGGTCAACCCGGCTGGGGCGGTCAGCCTGGATGGGGAGGTCAACCA
44. GGTGCCGGAAATAAGCCAGGAGCAGGGAATAAACCGGGAGCTGGAAACAAGCCTGGGGCGGGTAATAAACCA	554. GGTTATGGAGGGCAGCCAGGATACGGCGGACAGCCTGGGTATGGTGGTCAACCGGGATATGGGGGTCAACCA
45. GGTGGAAATATGGCTCCAGGCGGTAATATGGCCCCAGGAGGAAACATGGCACCCGGAGGGAATATGGCACCA	555. GGTGGTGGACAAGTCCCTGGAGGCGGTCAAGTACCCGGAGGAGGGCAAGTTCCTGGGGGAGGTCAGGTACCA
46. GGTTTTGGAAATGCTCCGGGTTTCGGTAACGCTCCTGGATTTGGTAATGCCCCTGGCTTTGGGAATGCACCA	556. GGTCACCACGGTCAGCCAGGACATCACGGACAACCGGGTCATCATGGGCAGCCTGGACACCATGGACAACCA
47. GGTGGAAATTCTGCTCCAGGCGGTAACTCAGCTCCGGGAGGTAATAGTGCCCCAGGAGGAAACAGTGCACCA	557. GGTGGGCAGCACATTCCTGGCGGTCAACATATACCCGGAGGTCAGCATATTCCAGGGGGACAACACATACCA
48. GGTGGTACGAATGCCCCTGGCGGTACAAATGCTCCCGGAGGTACTAACGCTCCTGGAGGAACTAATGCACCA	558. GGTCACCAAGGACTTCCTGGACATCAAGGTTTACCGGGGCATCAGGGTCTCCCAGGCCATCAAGGGTTACCA
49. GGTTGGGGCAATGCGCCGGGATGGGGAAACGCTCCTGGCTGGGGTAATGCTCCAGGGTGGGGGAATGCACCA	559. GGTCATCAAGGAAAGCCAGGACATCAGGGTAAACCGGGGCATCAAGGGAAGCCTGGACACCAAGGTAAACCA
50. GGTGGAAATTATGCCCCAGGCGGTAACTATGCTCCCGGAGGAAACTACGCTCCTGGAGGGAATTACGCACCA	560. GGTCATATGGGACAGCCTGGACATATGGGCCAACCCGGTCACATGGGTCAGCCAGGGCATATGGGGCAACCA
51. GGTGCGAATGGCGTACCCGGAGCTAATGGAGTCCCTGGGGCTAACGGTGTTCCAGGAGCAAATGGTGTACCA	561. GGTTTTCATGGTCAGCCAGGGTTCCATGGCAACCGGGATTCCACGGACAGCCTGGATTTCACGGCCAACCA
52. GGTGATGATGGAGCGCCAGGGGACGACGGTGCTCCCGGAGACGATGGGGCTCCTGGAGATGACGGCGCACCA	562. GGTGGGCATTCTCAACCCGGCGGTCATAGTCAGCCTGGAGGACATTCACAGCCAGGGGGACACTCGCAACCA
53. GGTCAGGGAGACGCTCCAGGCCAAGGAGATGCCCCAGGACAAGGTGATGCTCCGGGTCAAGGCGATGCACCA	563. GGTGGGCAGCACACTCCTGGCGGTCAACATACTCCCGGAGGTCAGCATACCCCTGGAGGACAACACACACCA
54. GGTGACGGTGAGGCTCCAGGAGATGGTGAAGCCCCTGGCGATGGAGAAGCTCCGGGTGATGGGGAAGCACCA	564. GGTGGTCATTGGCAACCCGGCGGTCACTGGCAACCTGGGGGACATTGGCAGCCTGGAGGGCATTGGCAACCA
55. GGTGGAGACGGTGCTCCCGGAGGAGATGGCGCTCCTGGAGGCGATGGAGCCCCTGGGGGTGATGGTGCACCA	565. GGTGGATACCATCAGCCAGGAGGTTATCATCAACCGGGCGGCTATCACCAGCCTGGGGGTTACCACCAACCA
56. GGTGGACATGATGCCCCAGGCGGCCATGACGCTCCTGGAGGTCACGATGCTCCCGGAGGACACGACGCACCA	566. GGTCACCAAGGCGTTCCAGGCCATCAAGGAGTCCCAGGACATCAAGGTGTACCGGGGCATCAGGGTGTACCA
57. GGTGCTGATGGAATACCCGGAGCCGATGGCATTCCTGGGGCGGACGGTATCCCTGGAGCAGATGGTATACCA	567. GGTATAGGACAAATTCCAGGAATCGGTCAGATACCCGGAATAGGTCAAATCCCTGGCATTGGGCAAATACCA
58. GGTGCTGACGGTTTGCCAGGGGCGGATGGTCTACCGGGAGCAGATGGACTTCCTGGAGCCGATGGGTTACCA	568. GGTGGCTTGCAGATTCCAGGAGGACTTCAAATTCCGGGCGGTTTACAAATCCCAGGGGGTCTACAGATACCA
59. GGTAAGGGAGACGCTCCAGGAAAAGGAGATGCTCCGGGTAAAGGCGATGCCCCTGGGAAAGGTGATGCACCA	569. GGTGGGCAGAAGATTCCAGGGGGACAAAAAATACCCGGAGGTCAGAAAATTCCTGGCGGTCAAAAGATACCA
60. GGTGGCATGGATGCCCCAGGAGGAATGGATGCTCCAGGCGGTATGGACGCTCCGGGAGGGATGGATGCACCA	570. GGTGGGCAAATGATCCCTGGCGGTCAAATGATACCCGGAGGTCAGATGATTCCTGGAGGACAAATGATACCA
61. GGTTTTGGCGATGCTCCTGGGTTCGGAGACGCTCCCGGCTTTGGTGATGCCCCTGGATTTGGAGATGCACCA	571. GGTATATTTGGACAGCCTGGAATCTTTGGTCAACCCGGAATTTTTGGGCAGCCAGGGATATTCGGTCAACCA
62. GGTGGGGCAGATTCTCCCGGTGGAGCCGATTCGCCAGGGGGTGCTGATAGTCCAGGAGGAGCGGATTCACCA	572. GGTGGCAGTCAGATACCGGGCGGATCACAAATTCCAGGGGGTTCTCAAATCCCAGGAGGCTCGCAAATACCA
63. GGTGCCGATGGAACTCCTGGAGCGGATGGTACGCCAGGGGCTGACGGTACTCCGGGAGCAGATGGGACACCA	573. GGTCAAGGAACTATCCCAGGACAGGGTACGATTCCAGGCCAAGGTACAATTCCGGGGCAAGGGACTATACCA
64. GGTGGGGATTGGGCCCCTGGCGGTGACTGGGCACCCGGAGGAGATTGGGCTCCAGGAGGTGATTGGGCACCA	574. GGTGGGCAATGGATACCCGGAGGTCAATGGATTCCTGGCGGTCAGTGGATCCCTGGAGGACAATGGATACCA
65. GGTGATGGTTATGCCCCTGGCGATGGATATGCTCCTGGAGACGGCTACGCTCCCGGAGATGGGTATGCACCA	575. GGTCAGGGGTACATTCCTGGGCAAGGCTACATACCGGGACAAGGTTATATTCCAGGCCAAGGATATATACCA
66. GGTGGTGATGCGGTTCCTGGAGGAGATGCAGTACCCGGAGGCGATGCTGTGCCAGGGGGAGACGCTGTACCA	576. GGTCAAGGTATTGTTCCAGGACAGGGTATAGTACCGGGGCAAGGAATCGTCCCAGGCCAAGGGATAGTACCA
67. GGTGGACAACAAGCCCCTGGCGGTCAACAGGCTCCTGGAGGTCAGCAAGCTCCCGGAGGACAGCAGGCACCA	577. GGTCAGTTAGGACTTCCTGGGCAACTCGGTCTCCCTGGACAATTAGGTTTACCCGGACAGCTTGGCTTACCA
68. GGTGAAGGAGCCCAACCCGGCGAAGGTGCTCAGCCTGGAGAAGGGGCACAGCCAGGAGAGGGAGCTCAACCA	578. GGTAAGCAAGGTTTACCCGGAAAACAAGGACTCCCAGGAAAGCAGGGTCTTCCAGGCAAACAGGGGTTACCA
69. GGTGGTGGACAAGCTCCTGGAGGAGGCCAAGCCCCAGGCGGAGGTCAAGCACCGGGAGGGGGTCAGGCACCA	579. GGTCAAGGAATGCTCCCAGGCCAAGGTATGCTTCCAGGACAAGGGATGTTACCGGGTCAGGGCATGTTACCA
70. GGTGGGCATCAGGCTCCTGGCGGTCATCAAGCCCCTGGAGGTCACCAAGCTCCCGGAGGACATCAAGCACCA	580. GGTTTTCTCGGACAGCCTGGATTCTTAGGTCAACCGGGATTTTTGGGGCAGCCAGGGTTCCTTGGTCAACCA
71. GGTATCGGTCAAGCGCCAGGAATAGGTCAGGCTCCGGGTATAGGACAAGCCCCTGGGATTGGACAGGCACCA	581. GGTCAAAGTGGACTCCCAGGACAATCTGGCTTACCCGGACAGAGCGGTCTTCCAGGGCAATCAGGCTTACCA
72. GGTGCCCAAGGTCTTCCAGGCGCTCAGGGTTTACCGGGAGCACAAGGACTTCCTGGGGCTCAAGGGTTACCA	582. GGTACGCTCGGACAGCCTGGAACATTAGGTCAACCCGGTACACTTGGGCAGCCAGGGACTCTAGGACAACCA
73. GGTGGAAAGCAGGCTCCTGGCGGTAAGCAAGCTCCCGGAGGTAAACAAGCCCCTGGAGGAAAACAGGCACCA	583. GGTGGTCTTTGGCAACCCGGCGGTTTATGGCAGCCTGGGGGATTGTGGCAACCTGGAGGACTCTGGCAACCA
74. GGTGGTCAGATGGCCCCAGGAGGACAAATGGCACCGGGAGGTCAAATGGCTCCAGGCGGACAGATGGCACCA	584. GGTCAGTACGGACTCCCAGGACAATATGGTCTTCCAGGGCAGTATGGCTTACCCGGTCAATACGGCTTACCA
75. GGTTTTGGACAAGCTCCCGGATTCGGTCAAGCGCCAGGGTTTGGTCAGGCTCCAGGATTTGGGCAAGCACCA	585. GGTCAGCTCGGAGTCCCAGGGCAGTTAGGTGTACCCGGTCAACTTGGCGTTCCAGGACAATTAGGCGTACCA
76. GGTGCCCAAGGTAGTCCAGGCGCTCAGGGTTCTCCTGGGGCTCAAGGAAGTCCCGGAGCACAAGGGTCACCA	586. GGTAAGAAGGGACAGCCAGGAAAAAAAGGTCAACCGGGTAAAAAGGGGCAGCCTGGAAAGAAAGGACAACCA
77. GGTGCCGGTCAAACCCCTGGGGCTGGACAAACTCCGGGAGCGGGTCAGACGCCAGGAGCAGGCCAAACACCA	587. GGTATGCAAGGAAAACCGGGCATGCAGGGTAAGCCAGGAATGCAAGGGAAGCCTGGGATGCAAGGTAAACCA
78. GGTTGGGGCCAAGCGCCGGGATGGGGTCAGGCTCCTGGCTGGGGACAAGCTCCAGGGTGGGGGCAAGCACCA	588. GGTGGGTTCCAAAAGCCAGGAGGTTTTCAGAAGCCTGGGGGATTCCAGAAACCGGGCGGATTTCAAAAACCA
79. GGTCAAGGTTATGCCCCAGGCCAAGGCTATGCTCCGGGTCAGGGATACGCTCCAGGACAAGGATATGCACCA	589. GGTCAAGGATCAAAGCCTGGACAAGGGAGTAAACCGGGGCAAGGTTCTAAGCCAGGACAGGGTAGTAAACCA
80. GGTCAAGGTGCCGTCCCAGGCCAAGGGGCTGTACCGGGTCAGGGAGCAGTTCCTGGACAAGGAGCGGTACCA	590. GGTGGGCAGAAGACTCCCGGAGGTCAGAAAACTCCTGGCGGTCAAAAAACCCCTGGAGGACAAAAGACACCA
81. GGTGGTGCCGAAGAGCCTGGGGGAGCAGAAGAACCCGGTGGAGCTGAGGAGCCAGGAGGGGCGGAAGAACCA	591. GGTGGAAAATGGCAGCCTGGAGGGAAATGGCAACCGGGCGGAAAGTGGCAACCTGGGGGTAAATGGCAACCA
82. GGTGGGGAAGGGGCTCCCGGAGGTGAGGGTGCTCCAGGAGGAGAAGGAGCGCCAGGCGGTGAAGGTGCACCA	592. GGTAAATACGGACAGCCTGGAAAATATGGTCAACCGGGGAAGTATGGGCAGCCAGGAAAGTACGGCCAACCA
83. GGTCATGGGGAAGCTCCTGGGCATGGCGAGGCTCCGGGTCACGGTGAAGCCCCTGGACATGGAGAAGCACCA	593. GGTCAAGGGAAAGTACCGGGGCAAGGAAAAGTCCCAGGCCAAGGTAAAGTTCCAGGACAGGGTAAGGTACCA
84. GGTGCTGAAGGTATTCCAGGAGCAGAAGGAATCCCTGGGGCGGAAGGCATACCGGGAGCCGAGGGTATACCA	594. GGTGGAATGATGCAGCCAGGCGGTATGATGCAACCCGGAGGGATGATGCAACCTGGGGGCATGATGCAACCA
85. GGTGCGGGTGAACTTCCTGGAGCCGGAGAATTGCCTGGGGCTGGCGAACTACCGGGAGCAGGGGAATTACCA	595. GGTGGATTTATGCAGCCTGGCGGTTTTATGCAACCCGGAGGATTCATGCAACCTGGGGGCTTTATGCAACCA
86. GGTGAAGGTAAAGCCCCTGGCGAAGGCAAAGCTCCTGGAGAGGGAAAAGCACCCGGAGAAGGAAAGGCACCA	596. GGTATGCAAGGGTCGCCCGGAATGCAAGGTTCTCCTGGGATGCAAGGAAGTCCAGGCATGCAGGGGTCACCA
87. GGTGCCGGTATGGAGCCAGGAGCAGGGATGGAACCCGGAGCTGGAATGGAACCTGGGGCGGGAATGGAACCA	597. GGTCAGATGGGCACTCCAGGACAAATGGGAACTCCGGGTCAAATGGGTACGCCAGGCCAAATGGGGACACCA
88. GGTGGGTTTGAGGCTCCTGGAGGTTTCGAAGCCCCTGGCGGTTTTGAAGCTCCCGGAGGATTTGAAGCACCA	598. GGTGGTATGTGGCAACCCGGCGGAATGTGGCAACCTGGAGGGATGTGGCAGCCTGGGGGCATGTGGCAACCA
89. GGTGGAGAAGCGAGCCCTGGCGGTGAAGCAAGTCCGGGTGGGGAAGCTAGTCCAGGAGGAGAGGCATCACCA	599. GGTTATATGGGGCAGCCTGGATATATGGGACAACCCGGATACATGGGTCAGCCAGGGTATATGGGCCAACCA
90. GGTGGAGCTACGGAACCCGGCGGTGCTACTGAGCCAGGAGGGGCAACAGAGCCTGGGGGAGCGACTGAACCA	600. GGTGTTATGGGTCAGCCAGGAGTCATGGGACAACCCGGCGTAATGGGGCAGCCTGGAGTGATGGGCCAACCA
91. GGTGGATGGGCGGAACCGGGAGGCTGGGCCGAGCCTGGAGGTTGGGCAGAGCCAGGAGGGTGGGCTGAACCA	601. GGTGGATTTTTCCAGCCAGGAGGGTTCTTTCAGCCTGGGGGATTCTTCCAACCGGGCGGCTTTTTCAACCA
92. GGTGGAGCTTACGAGCCTGGGGGAGCGTATGAACCCGGCGGTGCTTATGAGCCAGGAGGGGCATATGAACCA	602. GGTTCTTTTGGGCAGCCAGGATCATTTGGTCAACCCGGTAGCTTTGGACAGCCTGGAAGTTTCGGACAACCA
93. GGTGCCGGAGAAGTTCCTGGAGCGGGTGAAGTCCCAGGGGCTGGCGAAGTACCGGGAGCAGGAGAGGTACCA	603. GGTGGGCAGTTCACTCCTGGAGGACAATTTACCCCTGGCGGTCAGTTTACTCCCGGAGGTCAATTCACACCA
94. GGTGGCGGAGGGGCTCCTGGGGGTGGAGGTGCTCCAGGGGGAGGAGGCGCTCCCGGCGGTGGTGGGGCACCA	604. GGTGGATTCTGGCAACCTGGAGGATTTTGGCAACCGGGCGGCTTTTGGCAGCCTGGGGGTTTTTGGCAACCA
95. GGTGGTGGACATGCTCCCGGTGGGGGTCACGCTCCTGGCGGAGGTCATGCCCCTGGAGGAGGCCATGCACCA	605. GGTTTTTACGGACAGCCTGGATTCTATGGTCAACCGGGATTTTATGGGCAGCCAGGGTTCTACGGCCAACCA
96. GGTGGAGGAGCGATTCCAGGAGGGGGAGCAATACCGGGTGGTGGGGCTATCCCTGGCGGAGGTGCCATACCA	606. GGTTTTCAGGGCGTGCCAGGATTTCAAGGAGTCCCAGGCTTCCAAGGTGTTCCGGGTTTCCAGGGGGTACCA
97. GGTGGTGGACTTGCCCCTGGGGGTGGCTTAGCTCCCGGAGGAGGTCTCGCTCCTGGAGGCGGTTTAGCACCA	607. GGTTCACAAGGAAGTCCTGGATCGCAAGGTAGCCCAGGCAGTCAGGGGTCTCCGGGATCTCAAGGGTCACCA
98. GGTGGTGGGGCAAAGCCTGGAGGCGGAGCGAAACCGGGAGGAGGTGCTAAGCCAGGGGGTGGAGCTAAACCA	608. GGTGGGCAGTCTACTCCTGGAGGACAAAGTACCCCTGGCGGTCAATCAACTCCCGGAGGTCAGAGCACACCA
99. GGTGGTGGAATGGCACCCGGCGGAGGCATGGCCCCAGGGGGAGGGATGGCTCCTGGAGGAGGTATGGCACCA	609. GGTCAATGGGGCTCGCCAGGGCAATGGGGTTCTCCTGGACAATGGGGAAGCCCGGGTCAGTGGGGGTCACCA
100. GGTTTCGGAGGTGCTCCCGGTTTTGGGGGTGCCCCTGGATTTGGAGGCGCTCCTGGCTTTGGTGGTGCACCA	610. GGTTATCAGGGCAGTCCTGGATATCAAGGATCGCCTGGCTATCAAGGGTCTCCGGGATACCAAGGTTCACCA
101. GGTGCTGGGGGTAGTCCTGGAGCGGGTGGATCTCCGGGAGCCGGTGGTTCGCCAGGGGCAGGAGGCTCACCA	611. GGTCAAGGTTCTGTACCGGGGCAGGGTAGTGTTCCAGGCCAAGGCTCAGTCCCAGGACAAGGATCGGTACCA
102. GGTGGAGCAGGAACTCCTGGGGGTGCCGGTACGCCTGGAGGAGCGGGTACTCCCGGAGGGGCTGGCACACCA	612. GGTCAAGGTACTACGCCTGGACAAGGAACAACTCCTGGGCAAGGGACGACTCCGGGTCAGGGTACAACACCA
103. GGTTGGGGCGGTGCTCCAGGGTGGGGAGGGGCTCCTGGCTGGGGAGGCGCTCCGGGATGGGGTGGTGCACCA	613. GGTCAGTGGGGAACTCCAGGCCAATGGGGTACTCCGGGTCAATGGGGCACGCCAGGACAATGGGGGAACCA
104. GGTGGAGGCTATGCTCCGGGAGGAGGTTATGCCCCTGGCGGTGGTTACGCTCCTGGGGGTGGGTATGCACCA	614. GGTTATCAAGGTACTCCCGGATACCAAGGAACCCCAGGCTATCAGGGCACGCCAGGATATCAAGGGACACCA
105. GGTGGAGCCGGTGTCCCTGGGGGTGCTGGCGTTCCAGGAGGAGCGGGAGTACCCGGAGGGGCAGGCGTACCA	615. GGTGGTACTGTTCAGCCAGGAGGGACCGTACAGCCTGGCGGCACAGTCCAACCCGGCGGAACTGTCCAACCA
106. GGTGGACATCATGCCCCAGGCGGTCACCATGCTCCCGGAGGACACCACGCTCCAGGAGGTCATCACGCACCA	616. GGTGGATGGTGGCAGCCTGGAGGGTGGTGGCAACCTGGGGGTTGGTGGCAACCGGGCGGCTGGTGGCAACCA
107. GGTGCCCACGGTATTCCTGGAGCACATGGAATCCCTGGGGCTCATGGCATACCCGGAGCGCATGGTATACCA	617. GGTTATGGGGACAGCCAGGGTATTGGGGCCAGCCCGGATATTGGGGTCAGCCTGGATACTGGGGCCAACCA
108. GGTCACGGAGCTTTACCGGGTCATGGTGCTCTCCCAGGGCATGGCGCACTTCCAGGACATGGGGCTTTACCA	618. GGTGGTGTATGGCAACCTGGGGAGTCTGGAGGAGTGGCAACCCGGCGGAGTTTGGCAACCA
109. GGTGGGAAGCACGCACCCGGAGGTAAACACGCTCCAGGAGGAAAACATGCCCCTGGCGGTAAGCATGCACCA	619. GGTTATTACGGACAGCCTGGATATTATGGTCAACCGGGATACTATGGGCAGCCAGGGTACTACGGCCAACCA
110. GGTGGTCATATGGCCCCAGGAGGTCACATGGCTCCAGGCGGACATATGGCACCGGGAGGACACATGGCACCA	620. GGTGGGCAGTACGTTCCTGGCGGTCAGTATGTACCCGGAGGTCAATATGTTCCAGGGGGACAATACGTACCA
111. GGTTTTGGACATGCTCCAGGATTTGGTCACGCTCCGGGGTTCGGTCATGCGCCAGGCTTTGGGCATGCACCA	621. GGTGTTCAAGGAGTTCCTGGAGTCCAAGGGGTCCCAGGCGTACAGGGAGTGCCGGGAGTACAAGGTGTACCA
112. GGTCACGGAGCGTCTCCTGGACATGGTGCTAGTCCCGGTCATGGAGCTTCGCCAGGGCATGGGGCTTCACCA	622. GGTGGTGAGGAAGAACCGGAGGCGAAGAGGAACCTGGAGGTGAAGAAGAGCCAGGGGGAGAAGAAGAACCA
113. GGTCACGGAGCAACTCCCGGTCATGGAGCTACGCCAGGGCATGGTGCGACTCCTGGACATGGGGCTACACCA	623. GGTGGTGGCGAGGAGCCAGGAGGCGGAGAAGAACCCGGTGGAGGTGAAGAGCCTGGGGGAGGGGAAGAACCA
114. GGTTGGCATGGGGCTCCCGGCTGGCACGGTGCTCCTGGGTGGCATGGAGCGCCTGGATGGCATGGTGCACCA	624. GGTGAACACGGCGAGCCGGGGGAACATGGAGAACCCGGTGAGCACGGAGAGCCTGGGGAGCATGGTGAACCA
115. GGTTATGGACATGCTCCGGGGTATGGTCATGCGCCTGGCTACGGTCACGCTCCTGGATATGGGCATGCACCA	625. GGTGAAGGAGAGATCCCTGGGGAAGGCGAAATTCCGGGCGAGGGTGAGATTCCTGGAGAAGGGGAGATACCA
116. GGTGGTCATGTAGCCCCAGGCGGCCATGTTGCTCCCGGAGGACACGTCGCTCCTGGAGGGCATGTGGCACCA	626. GGTGAATTGGGCGAACCCGGTGAGTTAGGGGAGCCAGGAGAACTTGGAGAGCCTGGGGAATTAGGTGAACCA
117. GGTGCTATCGGTATCCCAGGAGCCATAGGAATACCCGGAGCAATAGGCATTCCAGGGGCGATTGGTATACCA	627. GGTGAAAGGGAGAACCGGGTGAGAAAGGAGAGCCTGGAGAGAAGGGCGAGCCAGGGGAAAAAGGTGAACCA
118. GGTGGACTAGCGATTCCGGGCGGTCTTGCTATCCCAGGGGGTTTGGCAATTCCAGGAGGCTTAGCTATACCA	628. GGTGGTATGGAAGACCTGGGGGCATGGAAGACCCGGCGGAATGGAGGAGCCAGGAGGGATGGAAGAACCA
119. GGTGGCAAAGCAATTCCAGGGGGTAAAGCGATTCCGGGCGGTAAGGCTATCCCAGGAGGAAAAGCTATACCA	629. GGTTTCGAAGGTGAGCCAGGATTTGAGGGCGAACCGGGTTTTGAAGGAGAGCCTGGATTCGAGGGAGAACCA
120. GGTGCTATGGGCATTCCTGGCGCAATGGGTATTCCGGGAGCCATGGGAATCCCTGGAGCGATGGGGATACCA	630. GGTGAATCTGGAGAGCCAGGGGAGTCAGGAGAACCCGGCGAAAGCGGTGAGCCTGGGGAAAGTGGTGAACCA

121. GGTTTTGGCATAGCTCCTGGGTTCGGAATTGCTCCCGGCTTTGGTATCGCCCCTGGATTTGGAATAGCACCA	631. GGTGAGGAAGGTACTCCTGGAGAAGAGGGAACCCCAGGCGAGGAGGGCACTCCGGGTGAAGAAGGAACACCA
122. GGTTCTGGGGCGATTCCAGGCTCAGGAGCAATACCCGGAAGTGGAGCTATCCCTGGCAGCGGTGCGATACCA	632. GGTGAATGGGGCGAGCCTGGAGAGTGGGGAGAACCCGGAGAATGGGGTGAGCCAGGGGAATGGGGGGAACCA
123. GGTGGTACGATTGCTCCGGGAGGTACAATAGCCCCAGGCGGTACTATCGCTCCTGGAGGAACTATAGCACCA	633. GGTTATGAAGGTGAGCCAGGATACGAGGGAGAGCCTGGATATGAGGGCGAACCGGGGTACGAAGGAGAACCA
124. GGTGCGGGGTGGATTCCAGGGGCAGGATGGATACCCGGAGCCGGTTGGATCCCAGGCGCTGGCTGGATACCA	634. GGTGGAGAGGAAGTACCGGGCGGAGAAGAGGTTCCAGGAGGGGAAGAAGTGCCTGGGGGTGAGGAGGTACCA
125. GGTGCTTATGGAATACCCGGAGCCTATGGCATTCCTGGAGCTTACGGTATCCCTGGGGCGTATGGTATACCA	635. GGTGGGGGAGGAGAGCCTGGAGGCGGAGGTGAGCCAGGAGGGGGTGGAGAACCCGGCGGTGGTGGCGAACCA
126. GGTGCCATAGGAGTTCCTGGAGCAATAGGCGTACCCGGAGCGATTGGTGTCCCTGGGGCTATCGGTGTACCA	636. GGTGGTGGACATGAGCCTGGAGGAGGGCATGAACCCGGAGGCGGTCACGAGCCAGGGGGAGGTCATGAACCA
127. GGTTTAGGTCTTGCTCCTGGATTAGGGTTAGCTCCCGGACTTGGTCTCGCCCCAGGCTTAGGATTGGCACCA	637. GGTATTGGTGGAGAGCCAGGAATCGGCGGAGAACCGGGCATAGGGGGTGAGCCTGGGATAGGAGGTGAACCA
128. GGTAAAGGACTCGCCCCTGGCAAAGGTCTTGCTCCTGGAAAAGGCTTAGCTCCGGGTAAGGGATTAGCACCA	638. GGTGGGGAAGGTCTTCCTGGAGGAGAGGGCTTGCCAGGCGGAGAAGGACTACCCGGTGGTGAAGGGTTACCA
129. GGTGGGATGCTTGCACCCGGAGGAATGTTAGCCCCTGGCGGTATGCTCGCTCCTGGAGGCATGTTGGCACCA	639. GGTGGGGAAGGTAAGCCAGGAGGAGAAGGAAAACCCGGCGGAGAGGGCAAGCCTGGAGGTGAAGGGAAACCA
130. GGTTTTGGGGCGTTGCCTGGCTTTGGTGCATTACCGGGATTTGGAGCACTTCCAGGGTTCGGAGCTTTACCA	640. GGTGGTATGGGAGAACCCGGAGGAATGGGTGAGCCAGGAGGCATGGGCGAGCCTGGAGGGATGGGGGAACCA
131. GGTGGATCTGCTCTTCCTGGGGGATCAGCTTTGCCTGGCGGTAGTGCATTACCCGGAGGTTCGGCTTTACCA	641. GGTGGTGGATTTGAGCCTGGAGGCGGATTCGAGCCAGGGGGTGGGTTTGAACCGGGAGGAGGTTTTGAACCA
132. GGTGGTGCGACTCTTCCTGGAGGAGCAACATTACCGGGAGGGGCTACACTGCCTGGCGGAGCTACGTTACCA	642. GGTGGTGGAAGTGAGCCTGGAGGAGGTAGCGAACCGGGCGGAGGGTCAGAGCCAGGGGGAGGCAGTGAACCA
133. GGTTGGGGCTTAGCTCCGGGATGGGGACTTGCTCCTGGCTGGGGTCTCGCTCCAGGGTGGGGGTTAGCACCA	643. GGTGAAGGTGGCACTCCAGGAGAAGGCGGTACACCCGGAGAGGGAGGAACCCCTGGGGAAGGAGGTACACCA
134. GGTTATGGTCTCGCCCCTGGGTATGGACTTGCTCCCGGATATGGCTTAGCTCCAGGATACGGTTTAGCACCA	644. GGTTGGGGAGGTGAGCCTGGATGGGGCGGAGAACCCGGCTGGGGAGGAGAGCCAGGGTGGGGTGGTGAACCA
135. GGTGCATTAGGAGTTCCTGGGGCTCTCGGTGTCCCTGGAGCACTTGGCGTACCCGGAGCTTTAGGCGTACCA	645. GGTTATGGGGGCGAGCCTGGATACGGAGGAGAACCGGGATATGGAGGTGAGCCAGGGTATGGTGGTGAACCA
136. GGTGGAGCTAAGAAGCCTGGGGGAGCGAAAAAACCCGGCGGTGCTAAAAAGCCAGGAGGGGCAAAAAAACCA	646. GGTGGGGAAGGTGTTCCTGGAGGAGAAGGAGTACCGGGCGGAGAGGGAGTGCCAGGGGGTGAAGGCGTACCA
137. GGTGGAAAAATGGCCCCAGGCGGTAAAATGGCACCCGGAGGAAAGATGGCTCCTGGAGGGAAGATGGCACCA	647. GGTGGACATCACGAGCCTGGGGGACACCATGAGCCAGGAGGGCACCACGAACCCGGCGGTCATCATGAACCA
138. GGTGGAGCTTTTAAGCCAGGAGGGGCATTCAAACCCGGCGGTGCGTTCAAGCCTGGGGGAGCATTTAAACCA	648. GGTGAGCACGGAATCCCTGGAGAACATGGAATACCCGGCGAGCATGGCATTCCTGGGGAACACGGTATACCA
139. GGTGGGTCTAAAGCTCCCGGAGGTTCAAAAGCCCCTGGCGGTAGTAAGGCTCCTGGAGGAAGTAAAGCACCA	649. GGTCACGAGGGTCTTCCTGGGCATGAAGGTTTACCGGGACACGAAGGACTTCCAGGCCATGAGGGGTTACCA
140. GGTAAGGGAACCGCTCCTGGCAAAGGCACAGCCCCTGGAAAAGGAACTGCTCCGGGTAAAGGTACGGCACCA	650. GGTGGGGAACACAAGCCTGGAGGAGAGCATAAACCGGGCGGAGAACATAAGCCAGGGGGTGAGCACAAACCA
141. GGTGGTGCTTGGAAGCCAGGAGGGGCATGGAAACCTGGGGGAGCCTGGAAACCCGGCGGAGCGTGGAAACCA	651. GGTCATGGGATGGAACCCGGACATGGTATGGAGCCAGGGCATGGAATGGAACCTGGACACGGTATGGAACCA
142. GGTGCCGGATATAAGCCAGGAGCAGGGTATAAACCGGGAGCTGGATACAAGCCTGGGGCGGGTTATAAACCA	652. GGTCATTTCGGCGAGCCAGGACATTTTGGAGAACCCGGACACTTTGGTGAGCCTGGGCACTTCGGTGAACCA
143. GGTGTAGGGGCAAAGCCTGGAGTGGGAGCGAAACCCGGCGTTGGTGCTAAGCCAGGAGTCGGAGCTAAACCA	653. GGTCACGAAGGATCGCCTGGGCATGAAGGTAGTCCTGGACACGAGGGTTCTCCCGGACATGAGGGGTCACCA
144. GGTGGTATGATGGCACCCGGAGGAATGATGGCCCCAGGCGGCATGATGGCTCCTGGAGGGATGATGGCACCA	654. GGTCACGAAGGTACTCCTGGACATGAAGGAACTCCGGGACACGAGGGTACGCCTGGGCATGAGGGGACACCA
145. GGTGGGATGTTTGCCCCTGGAGGAATGTTCGCTCCTGGCGGTATGTTTGCTCCGGGAGGCATGTTTGCACCA	655. GGTTGGCATGGAGAGCCTGGATGGCACGGCGAGCCAGGGTGGCATGGGGAACCGGGCTGGCATGGTGAACCA
146. GGTGGAATGGCGTCTCCGGGAGGCATGGCTAGTCCAGGAGGGATGGCAAGCCCTGGGGGTATGGCCTCACCA	656. GGTCATGGTTATGAGCCAGGGCATGGGTATGAACCGGGACATGGATACGAGCCTGGACACGGATATGAACCA
147. GGTGCCATGGGTACTCCCGGAGCGATGGGAACTCCAGGGGCTATGGGGACGCCAGGAGCAATGGGCACACCA	657. GGTGGTGAACACGTTCCTGGAGGAGAGCACGTCCCTGGGGGTGAGCATGTACCGGGCGGAGAACATGTACCA
148. GGTGGAATGTGGGCTCCTGGAGGGATGTGGGCACCCGGAGGCATGTGGGCCCCTGGCGGTATGTGGGCACCA	658. GGTATAGGAGAAATCCCTGGAATCGGTGAAATACCCGGAATAGGCGAAATTCCTGGCATTGGGGAAATACCA
149. GGTATGTATGGAGCACCCGGCATGTACGGAGCCCCTGGGATGTATGGCGCTCCTGGAATGTATGGTGCACCA	659. GGTGAGCTAGGAATCCCAGGGGAACTCGGTATTCCAGGCGAACTTGGCATACCCGGAGAATTAGGTATACCA
150. GGTATGGGCGTAGCTCCAGGGATGGGGGTTGCTCCTGGCATGGGTGTCGCTCCGGGAATGGGAGTAGCACCA	660. GGTGGCAAGGAAATTCCTGGGGGTAAGGAGATACCGGGCGGAAAAGAAATCCCTGGAGGTAAAGAGATACCA
151. GGTTTCGGATTCGCTCCCGGCTTTGGATTTGCTCCTGGCTTCGGTTTTGCCCCAGGGTTTGGGTTCGCACCA	661. GGTGAAGGTATGATCCCAGGAGAAGGAATGATACCCGGAGAGGGCATGATTCCAGGCGAAGGGATGATACCA
152. GGTGCTTTTGGATCGCCAGGAGCCTTCGGTAGTCCGGGAGCGTTTGGTTCTCCAGGGGCATTTGGGTCACCA	662. GGTGAGGGCTTTATTCCAGGCGAAGGGTTCATTCCTGGGGAAGGTTTTATACCCGGAGAAGGATTTATACCA
153. GGTTTCGGCACAGCTCCGGGTTTTGGGACCGCCCCTGGATTTGGAACTGCTCCTGGCTTTGGTACGGCACCA	663. GGTGAGGGCAGTATTCCAGGGGAAGGATCTATACCCGGAGAAGGTTCAATTCCTGGCGAAGGCTCTATACCA
154. GGTTGGGGATTTGCTCCAGGGTGGGGGTTCGCTCCCGGATGGGGCTTTGCGCCTGGCTGGGGTTTTGCACCA	664. GGTGGAACCGAAATCCCTGGAGGTACAGAAATTCCGGGCGGCACAGAGATTCCTGGGGGTACTGAAATACCA
155. GGTGGCTTTTACGCTCCAGGAGGGTTCTATGCTCCCGGAGGATTTTATGCCCCAGGCGGTTTCTACGCACCA	665. GGTGAAGGATGGATTCCAGGAGAAGGTTGGATCCCAGGGGAAGGGTGGATACCGGGCGAAGGCTGGATACCA
156. GGTGCTTTTGGAGTCCCAGGGGCATTTGGCGTACCCGGAGCCTTTGGTGTTCCAGGAGCTTTCGGTGTACCA	666. GGTGGTGAGTATATCCCTGGAGGAGAGTACATACCGGGCGGAGAATACATTCCTGGGGGTGAATATATACCA
157. GGTAGTGGAGCAAGTCCAGGGTCTGGGGCTTCGCCTGGCAGCGGTGCATCTCCCGGATCAGGAGCTTCACCA	667. GGTGTTGAAGGCATTCCGGGCGTAGAAGGAATCCCAGGGGTCGAAGGTATTCCAGGAGTAGAGGGTATACCA
158. GGTGGCACTGCTTCTCCTGGAGGAACCGCATCGCCTGGGGGTACAGCTAGTCCGGGAGGTACGGCTTCACCA	668. GGTGAGCTAGGTTTGCCAGGCGAACTTGGACTACCCGGAGAACTCGGTCTTCCTGGGGAATTAGGGTTACCA
159. GGTGGGTGGTCTGCTCCTGGCGGTTGGTCAGCTCCGGGAGGCTGGAGTGCCCCAGGGGGATGGTCGGCACCA	669. GGTTTGAAAGGAGAGCCAGGATTAAAAGGCGAACCCGGTCTTAAAGGTGAGCCTGGACTGAAGGGAGAACCA
160. GGTGGGTATAGTGCCCCTGGCGGTTACTCAGCTCCCGGAGGATATTCTGCTCCAGGAGGCTACAGTGCACCA	670. GGTATGCTTGGTGAGCCAGGGATGCTAGGAGAACCCGGCATGTTGGGCGAGCCTGGAATGTTAGGTGAACCA
161. GGTGGATCTGCGGTCCCAGGAGGAAGTGCTGTTCCAGGGGGCTCGGCAGTACCCGGAGGTTCAGCTGTACCA	671. GGTGGCTTTGAATTGCCTGGAGGGTTTGAACTTCCTGGGGGATTTGAGTTACCGGGCGGATTCGAGTTACCA
162. GGTGGTACAGCAACTCCGGGAGGCACGGCTACGCCAGGAGGTACTGCGACCCCTGGAGGAACCGCTACACCA	672. GGTGAATTAGGCAGTCCTGGCGAACTCGGTAGCCCAGGGGAACTTGGATCACCCGGAGAACTAGGTTCACCA
163. GGTGGATGGGCGACTCCGGGAGGGTGGGCTACCCCTGGAGGCTGGGCAACTCCTGGCGGTTGGGCCACACCA	673. GGTCTTGAAGGAACCCCAGGATTGGAAGGTACTCCCGGACTAGAGGGCACGCCTGGCTTAGAAGGGACACCA
164. GGTTATGCTGGCACTCCCGGATATGCAGGAACCCCAGGCTACGCCGGTACTCCAGGGTATGCGGGCACACCA	674. GGTTGGGAAGGTCTTCCAGGATGGGAGGGATTGCCAGGCTGGGAAGGACTACCGGGGTGGGAAGGGTTACCA
165. GGTGCTGTAGGAACTCCTGGAGCCGTGGGTACTCCCGGAGCGGTTGGTACGCCAGGGGCAGTAGGCACACCA	675. GGTTATCTCGGAGAGCCAGGGTATCTAGGTGAGCCTGGATATTTGGGCGAACCGGGATACTTAGGAGAACCA
166. GGTGGCTGGTGGGCTCCTGGAGGTTGGTGGGCCCCAGGCGGATGGTGGGCACCGGGAGGGTGGTGGGCACCA	676. GGTCTCGGAGTTGAGCCTGGACTAGGGGTAGAACCGGGATTAGGTGTAGAGCCAGGATTGGGCGTAGAACCA
167. GGTTATGGTTGGGCTCCAGGATACGGCTGGGCACCGGGATATGGGTGGGCCCCTGGCTATGGATGGGCACCA	677. GGTGAGAAAGGTAAGCCAGGAGAGAAGGGTAAACCCGGCGAAAAGGGCAAGCCTGGGGAAAAAGGAAAACCA
168. GGTGGATGGGTAGCTCCAGGAGGCTGGGTCGCTCCCGGAGGGTGGGTTGCCCCTGGCGGTTGGGTGGCACCA	678. GGTGGTAAAATGGAGCCAGGAGGGAAGATGGAACCCGGCGGAAAAATGGAACCTGGGGGCAAAATGGAACCA
169. GGTTATGGGTACGCTCCTGGATACGGATATGCCCCAGGGTATGGCTATGCTCCCGGTTACGGCTACGCACCA	679. GGTGGGTTTGAAAAGCCTGGAGGTTTTGAGAAACCGGGCGGATTCGAGAAGCCAGGGGGATTTGAAAAACCA
170. GGTGTAGGCTATGCGCCAGGAGTAGGATATGCCCCTGGAGTCGGTTACGCTCCCGGAGTTGGGTATGCACCA	680. GGTGGAAAAGAAAGTCCTGGCGGTAAAGAGAGCCCTGGGGGTAAGGAATCTCCCGGAGGCAAGGAGTCACCA
171. GGTGTCGGAGTAGCCCCAGGGGTAGGAGTCGCTCCCGGTGTTGGCGTTGCTCCTGGAGTGGGTGTAGCACCA	681. GGTGGAAAAGAGACTCCTGGCGGTAAGGAAACTCCGGGAGGTAAAGAAACCCCTGGAGGCAAGGAGACACCA
172. GGTAGAAGAGGACGCCCAGGAAGGAGGGGTAGGCCAGGGAGACGTGGAAGACCCGGCAGACGAGGTCGACCA	682. GGTAAATGGGGAGAGCCAGGAAAATGGGGCGAACCCGGCAAATGGGGTGAGCCTGGAAAGTGGGGGGAACCA
173. GGTGGGAGAAACAGGCCAGGGGGCAGAAATAGACCCGGCGGAAGGAATCGTCCTGGAGGTAGGAACCGACCA	683. GGTTACAAAGGTGAGCCAGGATACAAGGGCGAACCGGGGTATAAAGGAGAGCCTGGATATAAGGGAGAACCA
174. GGTGGGCGTGATCGTCCTGGAGGTAGAGATAGGCCAGGGGGAAGAGACAGACCCGGAGGAAGGGATCGACCA	684. GGTGAGGGTAAAGTTCCAGGAGAAGGAAAAGTACCGGGCGAAGGCAAAGTCCCAGGGGAAGGTAAGGTACCA
175. GGTCAAGGGAGAAGACCGGGTCAGGGTAGGCGTCCTGGGCAAGGTAGACGGCCTGGACAAGGAAGGCGACCA	685. GGTGGTATGATGGAACCTGGGGGCATGATGGAACCCGGCGGAATGATGGAGCCAGGAGGGATGATGGAACCA
176. GGTCGTGGTAGAGAGCCAGGAAGGGGAAGAGAACCCGGCAGAGGACGTGAGCCTGGGCGAGGTAGGGAACCA	686. GGTGGAATGTTCGAGCCAGGAGGGATGTTTGAACCCGGCGGTATGTTTGAGCCTGGGGGCATGTTTGAACCA
177. GGTGGTGGAAGAAGACCGGGAGGGGGCCGTCGTCCTGGAGGAGGCAGAAGGCCAGGCGGAGGTAGACGACCA	687. GGTATGGAAGGATCTCCTGGGATGGAGGGTAGCCCCGGAATGGAAGGTTCGCCAGGCATGGAAGGGTCACCA
178. GGTGGGCGTCATAGGCCAGGGGGAAGACACAGACCCGGAGGAAGGCATCGTCCAGGAGGTAGACATCGACCA	688. GGTACAGGGATGGAGCCAGGAACAGGCATGGAACCCGGTACTGGAATGGAACCTGGGACTGGTATGGAACCA
179. GGTCGTAGGGGCATTCCAGGAAGGCGAGGAATCCCAGGGAGAAGAGGTATTCCGGGTAGACGTGGCATACCA	689. GGTGGTTGGATGGAGCCAGGAGGGTGGATGGAACCGGGCGGCTGGATGGAACCTGGGGGATGGATGGAACCA
180. GGTGGGCTTAGAAGACCCGGCGGATTGAGAAGGCCAGGAGGTTTACGTCGTCCAGGGGGACTTAGGCGACCA	690. GGTATGGGCTATGAACCGGGCATGGGATATGAGCCAGGGATGGGGTACGAGCCTGGAATGGGTTATGAACCA
181. GGTAAACGTGGTAGGCCAGGAAAAAGGGGAAGACCCGGCAAGCGAGGACGCCCTGGAAAGAGAGGTCGACCA	691. GGTGGTGTAATGGAGCCAGGAGGGGTCATGGAACCGGGCGGAGTTATGGAACCTGGGGGAGTGATGGAACCA
182. GGTCGTGGGATGCGTCCTGGAAGAGGTATGAGACCCGGTAGGGGCATGAGGCCAGGGAGAGGAATGCGACCA	692. GGTTTCTTCGGAGAGCCTGGGTTTTTCGGCGAGCCAGGATTCTTTGGTGAACCGGGATTTTTTGGAGAACCA
183. GGTTTCCGAGGACGCCCTGGATTCAGAGGTAGACCCGGCTTTAGGGGAAGGCCAGGATTTCGTGGGCGACCA	693. GGTGAGTTCGGTAGCCCAGGGGAATTCGGATCACCCGGCGAGTTTGGCAGTCCAGGAGAATTTGGTTCACCA
184. GGTCGTAGAGGATCTCCAGGAAGAAGGGGTAGTCCCGGTAGGCGAGGTAGCCCTGGCAGACGTGGGTCACCA	694. GGTGGATTCGAGACTCCTGGAGGATTTGAGACCCCTGGCGGTTTTGAAACTCCCGGAGGCTTTGAAACACCA
185. GGTACTAGGGGAAGGCCAGGAACTCGTGGTAGACCCGGCACTAGAGGACGCCCTGGAACAAGAGGTCGACCA	695. GGTTGGGGGTTCGAGCCTGGATGGGGATTTGAGCCAGGGTGGGGTTTTGAACCCGGCTGGGGCTTTGAACCA
186. GGTAGAGGTTGGAGGCCAGGCAGAGGGTGGAGACCGGGAAGAGGATGGCGTCCTGGAAGGGGCTGGCGACCA	696. GGTTATTTTGGTGAGCCAGGATACTTTGGAGAACCGGGATATTTCGGCGAGCCTGGGTACTTCGGAGAACCA
187. GGTGGGAGATATAGACCCGGCGGTAGGTATCGTCCTGGAGGAAGGTACAGGCCAGGGGGCAGATACCGACCA	697. GGTGGGGAATTCGTGCCAGGGGGTGAATTTGTACCGGGCGGAGAGTTTGTTCCTGGAGGAGAATTTGTACCA
188. GGTAGACGTGGAGTCCCAGGCAGAAGAGGTGTACCCGGTCGTAGGGGCGTTCCAGGAAGGAGAGGAGTACCA	698. GGTTCCGAAGGCAGCCCAGGAAGTGAAGGTAGTCCAGGGTCAGAAGGATCTCCCGGTAGCGAGGGGTCACCA
189. GGTGGGAACAATCGTCCTGGAGGTAACAACAGACCCGGCGGAAATAATAGGCCAGGGGGCAATAACCGACCA	699. GGTGGATCAGAAACCCCTGGAGGCAGTGAAACTCCCGGAGGAAGTGAGACTCCTGGCGGTAGCGAAACACCA
190. GGTGGGGATAATCGTCCTGGAGGAGATAACAGACCGGGCGGAGACAATAGGCCAGGGGGTGACAACCGACCA	700. GGTGAAGGGTGGTCTCCTGGGGAAGGTTGGAGCCCTGGCGAAGGATGGAGTCCCGGAGAAGGCTGGTCACCA
191. GGTAACAGAGGTCAGCCTGGGAATAGAGGACAACCCGGAAACCGAGGGCAGCCAGGAAATCGTGGACAACCA	701. GGTTACGAAGGATCGCCAGGCTATGAGGGTTCTCCCGGATATGAAGGTAGTCCTGGATACGAGGGGTCACCA
192. GGTAACAGAGGTGAGCCAGGAAATCGTGGAGAACCCGGCAATAGGGGCGAGCCTGGGAATAGAGGAGAACCA	702. GGTGTGGAAGGCAGCCCAGGGGTTGAAGGATCTCCCGGAGTAGAAGGTAGTCCTGGCGTAGAGGGTTCACCA
193. GGTGGTGGGAATAGGCCAGGCGGAGGAAATCGTCCAGGAGGGGGCAACAGACCCGGTGGAGGTAATCGACCA	703. GGTGGGACCGAAACCCCTGGAGGAACAGAGACTCCTGGCGGCACAGAAACTCCGGGAGGTACTGAAACACCA
194. GGTCATGGAAACAGACCCGGTCACGGAAATAGGCCAGGGCATGGAATCGTCCAGGACATGGTAATCGACCA	704. GGTTGGGAAGGCACTCCAGGATGGGAAGGAACGCCAGGGTGGGAAGGTACTCCGGGCTGGGAGGGGACACCA
195. GGTAACCGAGGAATCCCAGGCAATAGGGGAATACCCGGAAATCGTGGCATTCCAGGAAACAGAGGTATACCA	705. GGTGGGTACACTGAGCCAGGGGGATATACAGAGCCTGGAGGTTATACGGAACCGGGCGGCTATACTGAACCA
196. GGTGGACGCAATCTCCCTGGGGGCAGAAACTTACCCGGAGGTCGTAATCTTCCTGGCGGTAGAAATTTACCA	706. GGTGGTGAAACGGTTCCTGGAGGAGAAACAGTACCCGGAGGCGAAACTGTGCCAGGGGGAGAGACTGTACCA
197. GGTAATAAGGGGCGTCCTGGAAATAAAGGAAGACCGGGAAACAAGGGTAGGCCAGGCAACAAAGGTCGACCA	707. GGTTGGGGTTGGGAGCCAGGGTGGGGATGGGAACCCGGCTGGGGGTGGGAACCTGGATGGGGCTGGGAACCA
198. GGTGGGATGAATAGGCCAGGAGGTATGAATCGTCCAGGGGGCATGAACAGACCCGGCGGAATGAATCGACCA	708. GGTTATTGGGGAGAGCCAGGGTATTGGGGCGAACCCGGATATTGGGGTGAGCCTGGATACTGGGGGGAACCA
199. GGTTTTGGGAACCGTCCTGGCTTTGGTAATCGCCCTGGATTCGGCAATAGACCCGGATTTGGAAATCGACCA	709. GGTTGGGAAGGTGTTCCAGGGTGGGAAGGAGTGCCAGGATGGGAAGGCGTACCGGGCTGGGAGGGGGTACCA
200. GGTCGTAATGGCAGCCCTGGACGCAATGGATCTCCGGGTAGAAATGGTTCCCCAGGAAGAAACGGTTCACCA	710. GGTGGGTACTACGAGCCTGGAGGTTATTATGAGCCAGGGGGATACTATGAACCGGGCGGATATTACGAACCA
201. GGTAATCGTGGTACTCCTGGAAATAGAGGAACGCCCGGAAACCGAGGGACCCCAGGCAACAGAGGTACACCA	711. GGTGGAGAATACGTTCCTGGCGGTGAGTATGTCCCTGGGGGTGAATATGTGCCCGGAGGAGAGTACGTACCA
202. GGTGGGAATTGGCGTCCAGGGGGCAACTGGAGACCCGGAGGTAATTGGAGGCCAGGAGGAAATTGGCGACCA	712. GGTGAAGTCGGTGTGCCAGGAGAGGTTGGAGTTCCAGGGGAAGTAGGCGTACCCGGAGAAGTGGGCGTACCA
203. GGTAATGGTTATCGTCCTGGAAACGGATACAGACCCGGCAATGGCTACCGCCCAGGGAATGGATATCGACCA	713. GGTGGTGGAGGTATTCCAGGGGGAGGAGGGATACCCGGTGGGGGTGGCATTCCCGGAGGCGGGGGCATACCA
204. GGTCGTAATGGAGTCCCTGGACGCAATGGGGTTCCGGGTAGAAATGGTGTGCCAGGAAGAAACGGTGTACCA	714. GGTGGGGGTGGCTTACCCGGAGGGGGAGGTCTTCCCGGTGGTGGAGGACTTCCAGGAGGCGGGGGCTTACCA
205. GGTGACGGCGACAGACCGGGCGATGGTGATCGTCCAGGGGATGGAGACAGGCCAGGAGATGGGGACCGACCA	715. GGTGGTGGCGGCAAGCCTGGAGGCGGAGGAAAACCCGGTGGGGGAGGTAAGCCAGGAGGGGGTGGAAAACCA
206. GGTGGAAGGGATCAGCCAGGAGGTAGAGATCAACCCGGCGGTCGTGACCAGCCTGGGGGCAGAGACCAACCA	716. GGTGGCGGAGGCAGTCCCGGCGGTGGTGGAAGCCCTGGGGGAGGAGGTTCTCCTGGAGGGGGTGGGTCACCA
207. GGTGAGGATGGTAGGCCAGGAGAAGATGGGCGTCCAGGGGAAGACGGCAGACCCGGAGAGGACGGTCGACCA	717. GGTGGAGGAGGGACTCCAGGGGGAGGCGGCACTCCCGGTGGCGGAGGTACTCCTGGGGGTGGTGGGACACCA
208. GGTGGTGGAGACAGACCGGGAGGAGGTGATCGTCCTGGAGGCGGAGATAGGCCAGGGGGTGGGGATCGACCA	718. GGTGGGGGTGGAGTGCCTGGAGGGGGAGGTGTCCCAGGAGGCGGAGGAGTTCCCGGCGGTGGTGGCGTACCA
209. GGTCATGGTGATCGTCCAGGGCATGGAGATAGACCCGGACATGGGGATAGGCCAGGACACGGCGATCGACCA	719. GGTGGTCATGGAATCCCTGGGGGTCACGGTATTCCTGGAGGGCATGGGATACCCGGCGGACATGGCATACCA
210. GGTAGGGGCGATATACCCGGTCGTGGAGACATCCCAGGAAGAGGTGATATTCCAGGCAGAGGAGATATACCA	720. GGTGGTCATGGTTTACCCGGCGGACACGGTCTTCCTGGGGGACATGGATTGCCTGGAGGGCATGGCTTACCA
211. GGTGATCTTGGAAGGCCAGGAGATTTAGGTAGACCCGGAGACTTAGGACGCCCTGGGGATCTCGGTCGACCA	721. GGTCATGGAGGAAAACCGGGACATGGTGGTAAGCCTGGACACGGCGGAAAGCCAGGGCATGGGGGTAAACCA
212. GGTGACAAGGGAAGACCCGGAGACAAAGGTAGGCCAGGGGATAAAGGACGCCCAGGAGATAAGGGTCGACCA	722. GGTGGACATGGCTCTCCAGGCGGTCATGGTAGTCCGGGAGGGCATGGAAGTCCTGGGGGTCACGGCTCACCA
213. GGTGGGATGGATCGTCCTGGAGGTATGGATAGGCCAGGGGGCATGGACAGACCCGGCGGAATGGATCGACCA	723. GGTCATGGTGGAACTCCTGGACATGGAGGTACTCCCGGTCACGGAGGAACCCCTGGGCATGGCGGCACACCA
214. GGTGGGTTCGATAGGCCAGGGGGATTCGACAGACCCGGGCGGCTTTGATCGTCCAGGAGGTTTTGACCGACCA	724. GGTCACGGAGGAGTCCCTGGGCATGGTGGGGTTCCAGGACATGGAGGTGTACCCGGTCATGGCGGCGTACCA
215. GGTAGAGATGGTTCTCCGGGTCGTGATGGATCGCCTGGCAGAGACGGTAGTCCTGGAAGGGATGGGTCACCA	725. GGTATAGGGGGTATTCCTGGAATTGGTGGCATTCCCGGTATCGGCGGAATCCCTGGGATAGGAGGCATACCA
216. GGTAGAGACGGAACCCCAGGCAGAGATGGTACGCCCGGTCGTGATGGAACTCCAGGAAGGGATGGCACACCA	726. GGTGGTCTTGGCATTCCAGGAGGACTCGGTATTCCGGGCGGCTTAGGAATCCCTGGGGGTTTAGGCATACCA
217. GGTGATTGGGGAAGACCCGGAGATTGGGGCCGTCCTGGAGACTGGGGTAGGCCAGGCGATTGGGGGCGACCA	727. GGTAAAGGCGGTATCCCAGGGAAAGGTGGCATTCCGGGCAAAGGGGGTATTCCAGGAAAAGGAGGTATACCA
218. GGTGGGTATGATAGGCCAGGGGGCTACGACAGACCCGGCGGATATGACCGTCCTGGAGGTTACGATCGACCA	728. GGTGGCATGGGGATACCGGGAGGGATGGGTATTCCTGGAGGTATGGGAATCCCAGGCGGAATGGGCATACCA
219. GGTCGTGGCGATGTCCCAGGAAGAGGAGATGTTCCGGGTAGAGGTGATGTGCCTGGACGAGGGGATGTACCA	729. GGTGGTTTTGGCATTCCTGGGGGCTTTGGAATACCCGGCGGATTCGGTATCCCTGGAGGATTTGGTATACCA
220. GGTCAGGGCCAGCGTCCTGGGCAGGGTCAAAGACCGGGACAGGGGCAACGTCCAGGCCAAGGACAACGACCA	730. GGTGGCATTGGAAGCCCTGGAGGAATAGGTTCTCCCGGCGGAATCGGTAGTCCTGGGGGTATAGGCTCACCA
221. GGTCAGAGAGGTGAGCCTGGGCAAAGAGGAGAACCGGGTCAACGAGGGGAGCCAGGACAGCGTGGAGAACCA	731. GGTGGTGGAACTATCCCTGGAGGAGGTACGATTCCTGGCGGAGGCACAATACCGGGAGGGGGTACTATACCA
222. GGTAGAGGGGGACAGCCAGGAAGAGGAGGTCAACCGGGTCGTGGTGGCAGCCTGGACGAGGCGGACAACCA	732. GGTGGTGGGTGGATTCCTGGAGGGGGATGGATACCGGGAGGAGGTTGGATCCCTGGCGGAGGCTGGATACCA
223. GGTGGGCATCGTCAGCCAGGGGGACATAGACAGCCTGGAGGACACAGACAACCCGGCGGTCATAGGCAACCA	733. GGTGGTGGATACATACCGGGCGGAGGCTATATCCCTGGGGGAGGTTATATTCCTGGAGGAGGGTACATACCA
224. GGTATCCAAGGGCGTCCTGGAATACAAGGAAGACCGGGGATTCAGGGTAGGCCAGGCATTCAAGGTCGACCA	734. GGTGGAATCGGAGTCCCTGGGGGCATAGGAGTTCCTGGAGGTATTGGTGTACCCGGCGGTATAGGCGTACCA
225. GGTGGGCTTCAACGTCCAGGAGGTTTACAAAGGCCAGGGGGATTGCAGAGACCCGGCGGACTACAGCGACCA	735. GGTTTAGGTGGTCTTCCAGGACTTGGCGGATTGCCTGGGTTAGGGGGATTACCGGGCTTAGGAGGTTTACCA
226. GGTCAAAGGGGTAAGCCAGGACAGCGTGGAAAACCGGGACAACGAGGAAAGCCTGGGCAGAGAGGTAAACCA	736. GGTGGGAAAGGTTTGCCAGGGGGTAAGGGTCTTCCTGGAGGTAAAGGACTACCGGGCGGAAAAGGCTTACCA
227. GGTATGCAAGGGCGTCCTGGAATGCAAGGAAGACCGGGGATGCAGGGTAGGCCAGGCATGCAAGGTCGACCA	737. GGTATGGGTGGTTTGCCAGGAATGGGAGGTCTTCCAGGCATGGGCGGACTACCGGGGATGGGAGGGTTACCA
228. GGTCAAGGTTTTCGCCCAGGACAGGGGTTCCGTCCAGGCCAAGGCTTTAGACCGGGACAAGGATTTCGACCA	738. GGTGGTGGATTTTTGCCAGGGGGAGGTTTTCTACCGGGCGGAGGCTTTCTTCCTGGAGGAGGGTTCTTACCA
229. GGTAGACAAGGAAGTCCAGGACGCCAAGGTTCCCCAGGCAGACAGGGCTCTCCGGGTCGTCAAGGGTCACCA	739. GGTGGATTAGGATCTCCTGGGGGTTTAGGCAGTCCCGGAGGTCTAGGTAGCCCTGGCGGTCTTGGTTCACCA
230. GGTAGAGGACAAACCCCAGGAAGAGGTCAGACTCCCGGTCGTGGTCAAACGCCTGGACGAGGGCAAACACCA	740. GGTGGAGGTCTTACTCCTGGGGGTGGTTTAACTCCCGGAGGAGGGTTGACTCCAGGCGGTGGCTTGACACCA
231. GGTCAATGGGGTAGGCCAGGACAATGGGGAAGACCCGGACAGTGGGGCCGTCCAGGGCAATGGGGGCGACCA	741. GGTGGTGGATGGTTGCCTGGAGGAGGGTGGCTTCCTGGGGGAGGTTGGTTACCGGGCGGAGGCTGGTTACCA
232. GGTTATGGTCGTCAGCCTGGATACGGCAGACAACCGGGGTATGGAAGACAGCCAGGATATGGGAGACAACCA	742. GGTGGTGGATATTTGCCAGGGGGAGGTTATCTACCGGGCGGAGGGTATCTTCCTGGAGGAGGCTATTTACCA
233. GGTCAGGTTGGTAGACCGGGTCAAGTGGGGCGTCCAGGACAAGTCGGAAGGCCAGGGCAAGTAGGTCGACCA	743. GGTGGTGTAGGACTACCGGGAGGAGTCGGCTTGCCTGGCGGAGTAGGTCTTCCTGGAGGGGTTGGGTTACCA
234. GGTGAAAGAGGCGAGCCAGGAGAGCGTGGGGAACCGGGTGAGAGAGGGGAGCCTGGAGAACGAGGTGAACCA	744. GGTGGAAAAGGAAAGCCTGGGGGTAAAGGTAAGCCAGGAGGGAAGGGGAAACCGGGCGGCAAAGGCAAACCA
235. GGTGAAGGGGGCCGTCCTGGAGAGGGAGGAAGACCCGGAGAAGGAGGTAGGCCAGGCGAAGGTGGGCGACCA	745. GGTGGTGGAATGAAACCTGGAGGAGGGATGAAGCCTGGGGGTGGCATGAAACCGGGCGGAGGTATGAAACCA
236. GGTCACGAGGGGCGTCCAGGGCATGAAGGAAGACCGGGACACGAAGGTAGGCCAGGACATGAGGGTCGACCA	746. GGTGGGTTTGGTAAACCGGGCGGATTCGGCAAGCCTGGAGGTTTTGGAAAGCCAGGGGGATTTGGGAAACCA
237. GGTATTGAAGGGCGTCCTGGAATCGAAGGAAGACCGGGGATAGAGGGTAGGCCAGGCATAGAAGGTCGACCA	747. GGTAGCGGAGGAAAACCGGGAAGTGGTGGAAAGCCTGGATCTGGGGGCAAGCCAGGGTCAGGAGGTAAACCA
238. GGTGAAAGGGGATTGCCTGGCGAACGAGGTCTACCCGGAGAAAGAGGACTTCCAGGAGAGCGTGGGTTACCA	748. GGTGGAAAAGGCACTCCTGGCGGTAAGGGTACTCCCGGAGGTAAAGGAACTCCAGGGGGCAAAGGTACACCA
239. GGTGAGAAGGGACGCCCAGGGGAAAAAGGTAGGCCAGGAGAGAAAGGAAGACCCGGAGAAAAGGGTCGACCA	749. GGTGGATGGGGCAAGCCTGGAGGTTGGGGAAAGCCAGGAGGGTGGGGTAAACCGGGCGGCTGGGGAAACCA
240. GGTGGGGAAATGCGTCCTGGAGGAGAGATGAGACCGGGCGGAGAAATGAGGCCAGGGGGTGAAATGCGACCA	750. GGTGGTGGATATAAGCCTGGAGGAGGTTACAAGCCAGGGGGAGGGTATAAACCCGGAGGCGGTTATAAACCA
241. GGTAGGGGTTTCGAGCCAGGACGAGGATTTGAACCGGGAAGAGGTTTTGAGCCTGGGCGTGGATTCGAACCA	751. GGTGGCAAAGGTGTCCCAGGGGGTAAGGGAGTACCGGGCGGAAAAGGAGTTCCTGGAGGTAAAGGCGTACCA
242. GGTGGACGAGAATCTCCTGGCGGTCGTGAAAGCCCTGGGGGTAGAGAAAGTCCCGGAGGCAGAGAGTCACCA	752. GGTGGAATGGGCTCCCCTGGAGGGATGGGAAGTCCCGGAGGCATGGGTTCTCCTGGCGGTATGGGGTCACCA
243. GGTGGAAGAACAGAGCCAGGAGGTAGAACGGAACCCGGCGGTCGTACTGAGCCTGGGGGACGCACTGAACCA	753. GGTGGAATGGGGACTCCCGGAGGCATGGGTACTCCTGGAGGGATGGGAACCCCTGGCGGTATGGGCACACCA
244. GGTGAAGGTTGGAGGCCAGGGGAAGGATGGAGACCCGGAGAAGGGTGGCGTCCAGGAGAGGGCTGGCGACCA	754. GGTATGGGTGGTGTACCGGGGATGGGAGGAGTTCCTGGAATGGGCGGAGTCCCTGGCATGGGAGGCGTACCA
245. GGTGGGGAATATAGGCCAGGGGGTGAGTACAGACCGGGCGGAGAGTATCGTCCTGGAGGAGAATACCGACCA	755. GGTGGATTTGGAAGTCCGGGAGGATTCGGTAGTCCTGGCGGCTTTGGTTCTCCAGGGGGTTTTGGGTCACCA
246. GGTGAAGTCGGCAGACCCGGAGAAGTAGGACGCCCAGGGGAAGTTGGAAGGCCAGGAGAGGTAGGTCGACCA	756. GGTGGATTCGGCACTCCTGGCGGTTTTGGAACCCCTGGAGGCTTTGGTACTCCGGGAGGATTTGGGACACCA
247. GGTGGTGGAGGTAGACCCGGTGGGGGAGGAAGGCCAGGAGGCGGGGGCCGTCCCGGAGGGGGTGGCCGACCA	757. GGTTTTGGCGGTGTGCCGGGATTTGGAGGAGTTCCAGGATTCGGAGGGGTCCCAGGCTTTGGTGGTGTACCA
248. GGTGGTGGACATAGGCCAGGGGGAGGTCACAGACCCGGAGGCGGTCATCGTCCTGGAGGAGGGCATCGACCA	758. GGTGGAGGTAGTAGCCCAGGGGGAGGGTCATCTCCTGGCGGTGGTTCAAGTCCCGGAGGAGGCAGTTCACCA

249. GGTGGTGGCATAAGACCGGGCGGAGGTATTCGTCCTGGAGGAGGGATACGGCCTGGGGGTGGAATCCGACCA	759. GGTAGTGGAGGAACCCCTGGCTCTGGGGGCACGCCTGGAAGCGGTGGTACTCCGGGATCAGGAGGTACACCA
250. GGTGGTGGATTACGGCCAGGGGGAGGCTTGAGACCGGGCGGAGGTTTGCGTCCTGGAGGAGGGTTACGACCA	760. GGTGGATGGGGTAGCCCTGGCGGTTGGGGATCTCCTGGAGGGTGGGGCAGTCCCGGAGGCTGGGGGTCACCA
251. GGTGGTGGCAAACGGCCTGGAGGAGGTAAAAGACCGGGCGGAGGGAAGCGTCCTGGGGGTGGAAAACGACCA	761. GGTTATGGCGGCAGCCCAGGATACGGTGGATCTCCGGGATATGGAGGAAGTCCAGGGTATGGGGGTTCACCA
252. GGTGGTGGGATGAGACCCGGTGGAGGTATGAGGCCAGGCGGAGGAATGCGTCCAGGAGGGGGCATGCGACCA	762. GGTAGCGGAGGTGTCCCTGGAAGTGGTGGAGTACCGGGTTCTGGGGGCGTGCCAGGCTCAGGAGGAGTACCA
253. GGTTTTGGAGGAAGGCCAGGATTTGGTGGTAGACCCGGCTTTGGGGGACGCCCTGGATTCGGAGGTCGACCA	763. GGTACAGGCGGTACTCCGGGTACGGGAGGAACTCCTGGCACAGGTGGAACCCCAGGGACTGGAGGTACACCA
254. GGTGGACGAGGTAGCCCAGGGGGTCGTGGTTCTCCGGGAGGAAGAGGCAGTCCTGGCGGTAGAGGGTCACCA	764. GGTGGTGGATGGACACCGGGCGGAGGCTGGACCCCTGGAGGAGGTTGGACTCCTGGGGGAGGGTGGACACCA
255. GGTGGTGGCACTCGTCCTGGAGGAGGTACACGGCCTGGGGGTGGAACAAGACCGGGCGGAGGGACTCGACCA	765. GGTGGATACGGCACTCCTGGCGGTTATGGAACCCCTGGAGGCTATGGTACTCCGGGAGGATATGGACACCA
256. GGTTGGGGAGGAAGACCCGGCTGGGGCGGTCGTCCTGGATGGGGTGGTAGGCCAGGGTGGGGAGGTCGACCA	766. GGTACAGGTGGTGTTCCTGGAACTGGCGGAGTACCGGGTACGGGAGGAGTCCCAGGCACAGGGGGCGTACCA
257. GGTTATGGTGGTAGGCCAGGATATGGGGGAAGACCCGGCTATGGAGGACGCCCTGGATACGGAGGTCGACCA	767. GGTTGGGGCGGTGTACCCGGATGGGGTGGAGTTCCAGGGTGGGGAGGAGTGCCAGGCTGGGGAGGTGTACCA
258. GGTGGTGGGGTACGGCCAGGAGGCGGAGTTCGTCCAGGGGGAGGAGTAAGACCGGGTGGAGGTGTACGACCA	768. GGTGGTGGATATGTTCCTGGAGGAGGTTATGTACCGGGCGGAGGGTATGTGCCAGGGGGAGGCTATGTACCA
259. GGTGGGCACCACAGACCCGGCGGACATCATAGGCCAGGGGGACACCATCGTCCAGGAGGTCATCACCGACCA	769. GGTGGTGGGGTTGTGCCAGGGGGAGGTGTAGTTCCAGGAGGCGGAGTCGTACCCGGTGGAGGAGTAGTACCA
260. GGTAGGGGACACATTCCAGGCAGAGGACATATACCCGGACGAGGTCATATTCCTGGGCGTGGTCACATACCA	770. GGTCATCATGGAATCCCTGGACATCACGGTATTCCTGGGCACCACGGCATACCCGGACACCATGGCATACCA
261. GGTGGGCATAGACTTCCTGGCGGTCACAGATTACCCGGAGGTCATCGTCTTCCAGGGGGACATAGGTTACCA	771. GGTCATCATGGACTCCCTGGGCACCACGGCTTACCCGGACATCACGGTCTTCCAGGACACCATGGCTTACCA
262. GGTCACAGAGGAAAGCCTGGGCATAGGGGAAAACCCGGACACCGAGGTAAGCCAGGACATCGTGGAAAACCA	772. GGTGGACACCATAAACCCGGCGGTCATCACAAGCCTGGGGGACATCATAAGCCAGGAGGGCACCACAAACCA
263. GGTCATATGGGTAGGCCAGGGCATATGGGAAGACCGGGTCACATGGGGCGTCCTGGACATATGGGCCGACCA	773. GGTCATGGTCACAGTCCGGGACATGGGCACTCTCCAGGCCACGGCCATTCGCCAGGGCATGGACACTCACCA
264. GGTTTTGGACATAGGCCAGGGTTCGGTCACAGACCCGGCTTTGGTCATCGTCCTGGATTTGGGCATCGACCA	774. GGTCACCACGGTACTCCGGGTCATCATGGGACTCCAGGGCACCATGGAACGCCAGGACATCACGGCACACCA
265. GGTCACGGTAGAAGTCCGGGTCATGGTCGTTCTCCAGGACATGGAAGATCGCCAGGGCATGGGAGGTCACCA	775. GGTCATCATGGAGTCCCTGGACATCACGGTGTTCCGGGTCACCATGGCGTTCCTGGGCACCACGGCGTACCA
266. GGTACGCATGGTAGGCCAGGGACACATGGGCGTCCAGGAACTCATGGAAGACCCGGCACTCACGGTCGACCA	776. GGTCACGGTATAATCCCTGGACATGGCATTATTCCGGGTCATGGGATAATTCCTGGGCATGGAATCATACCA
267. GGTTGGGGCCATAGACCGGGATGGGGTCATCGGCCTGGCTGGGGACATCGTCCAGGGTGGGGGCATCGACCA	777. GGTGGGCATATACTTCCAGGGGGACATATTCTACCCGGAGGTCACATTCTTCCTGGCGGTCATATCTTACCA
268. GGTTACCACGGACGCCCTGGATACCATGGAAGACCCGGCTATCATGGGAGGCCAGGATATCACGGTCGACCA	778. GGTGGCAAGCACATTCCAGGGGGTAAACATATTCCGGGCGGTAAGCATATCCCAGGAGGAAAACACATACCA
269. GGTCACCGAGGTGTTCCAGGACATCGTGGCGTCCCTGGGCACAGAGGAGTTCCGGGTCATAGAGGCGTACCA	779. GGTGGTCACATGATACCCGGAGGACATATGATCCCAGGCGGCCATATGATTCCAGGAGGGCATATGATACCA
270. GGTATCGGTATTCGTCCTGGAATTGGTATAAGACCCGGAATAGGCATACGCCCAGGGATAGGAATCCGACCA	780. GGTCATTTCGGAATCCCAGGGCACTTTGGAATACCCGGTCACTTCGGCATTCCTGGACATTTTGGCATACCA
271. GGTTTGAGAGGTATTCCAGGACTTAGGGGCATTCCGGGTCTACGAGGAATCCCAGGCTTGCGTGGCATACCA	781. GGTGGGCATTCAATTCCAGGGGGACACTCTATACCCGGAGGTCACAGTATTCCTGGCGGTCATAGCATACCA
272. GGTGGAAAACGCATTCCTGGGGGCAAGAGAATCCCTGGCGGTAAGCGAATACCCGGAGGTAAAAGGATACCA	782. GGTCACGGTACAATTCCTGGACATGGCACTATCCCTGGGCATGGTACGATTCCGGGTCATGGAACTATACCA
273. GGTGGGATAATGAGACCGGGCGGAATCATGCGTCCAGGGGGTATTATGAGGCCAGGAGGAATTATGCGACCA	783. GGTTGGCATGGGATTCCTGGGTGGCACGGTATCCCTGGATGGCATGGAATACCCGGCTGGCATGGTATACCA
274. GGTTTTGGACGCATTCCTGGATTTGGCAGAATACCGGGCTTTGGTCGTATCCCAGGGTTCGGAAGAATACCA	784. GGTGGCTACCACATTCCAGGAGGGTATCATATTCCGGGCGGTTACCATATCCCTGGGGGTTATCACATACCA
275. GGTCGTATAGGATCTCCGGGTAGAATCGGTTCGCCTGGCAGAATTGGGAGTCCTGGAAGAATAGGGTCACCA	785. GGTCATATTGGCGTACCCGGTCACATTGGTGTTCCTGGGCATATCGGAGTCCCTGGACATATAGGCGTACCA
276. GGTATAAGAGGAACCCCAGGCATTAGAGGGACGCCCGGTATCCGTGGTACTCCTGGAATACGAGGTACACCA	786. GGTGGACTTCACTTGCCAGGCGGATTACATTTACCGGGAGGGCTTCATCTTCCTGGGGGTCTACACTTACCA
277. GGTGGATGGAGGATACCCGGAGGTTGGAGAATCCCAGGGGGCTGGCGTATTCCAGGAGGGTGGCGAATACCA	787. GGTAAGGGGCATCTTCCAGGGAAAGGTCATCTACCGGGTAAAGGCCATCTCCCAGGAAAAGGACACTTACCA
278. GGTTATGGCAGAATCCCTGGATATGGACGAATACCGGGCTATGGTCGCATTCCTGGGTACGGAAGAATACCA	788. GGTCATATGGGACTTCCAGGACATATGGGCTTACCGGGTCACATGGGTTTGCCAGGGCATATGGGGTTACCA
279. GGTATAGGTGTAAGACCGGGAATAGGAGTACGCCCTGGAATCGGAGTTCGTCCTGGCATTGGGGTTCGACCA	789. GGTCATTTCGGACTCCCTGGGCACTTCGGCTTACCCGGTCACTTTGGTCTTCCAGGACATTTTGGCTTACCA
280. GGTTTACGAGGACTACCGGGATTGAGAGGTCTTCCAGGCTTGCGTGGGTTGCCAGGATTAAGGGGCTTACCA	790. GGTCACAGTGGTCTCCCAGGACATTCTGGCTTACCCGGTCATAGCGGACTTCCAGGGCATTCAGGCTTACCA
281. GGTAGAAAAGGTCTCCCTGGACGCAAAGGCTTACCCGGTCGTAAAGGATTGCCAGGAAGAAAGGGGTTACCA	791. GGTGGCTTGCATACTCCGGGAGGTTTACACACTCCAGGCGGTCTACATACCCCAGGAGGACTTCATACACCA
282. GGTCGTGGGATGCTTCCAGGAAGAGGTATGTTACCGGGTAGGGGAATGTTGCCAGGGCGAGGCATGTTACCA	792. GGTGGTCATTGGCTTCCTGGGGGACATTGGTTACCCGGCGGTCACTGGTTGCCTGGAGGGCATTGGTTACCA
283. GGTTTTAGGGGATTGCCTGGATTCCGAGGTCTTCCTGGGTTCAGAGGACTACCGGGCTTTCGTGGCTTACCA	793. GGTGGACATTACTTACCCGGAGGTCATTATCTTCCTGGGGGACACTATTTGCCTGGCGGTCACTACTTACCA
284. GGTGGAAGAAGTTTACCGGGCGGACGTTCACTCCCAGGGGGTAGATCATTGCCAGGAGGTCGTAGCTTACCA	794. GGTGGCTTACACGTTCCGGGCGGTTTGCATGTCCCTGGGGGTCTTCATGTGCCTGGAGGATTACATGTACCA
285. GGTCGTGGGCTAACTCCAGGCAGAGGACTTACTCCGGGTAGAGGTTTGACGCCAGGAAGGGGATTAACACCA	795. GGTAAGCATGGGAAGCCTGGAAAGCACGGAAAACCGGGTAAACACGGCAAGCCAGGGAAACATGGTAAACCA
286. GGTCGTTGGGGTTTACCGGGTAGATGGGGACTTCCAGGGAGGTGGGGCTTGCCAGGACGCTGGGGGTTACCA	796. GGTCATGGGATGAAACCTGGACATGGTATGAAGCCAGGGCATGGAATGAAACCCGGACACGGTATGAAACCA
287. GGTGGGCTCTACAGACCCGGCGGATTATATCGTCCAGGAGGTTTGTATAGGCCAGGGGGACTTTATCGACCA	797. GGTGGACATTTCAAGCCTGGGGGACACTTCAAACCCGGCGGTCATTTTAAGCCAGGAGGGCACTTTAAACCA
288. GGTGGAAGAGTTCTCCCAGGGGGTAGAGTATTGCCAGGAGGTCGTGTTTTACCGGGCGGACGAGTCTTACCA	798. GGTGGCAAACACAGCCCAGGCGGTAAACATTCTCCAGGGGGTAAGCATAGTCCCGGAGGAAAGCACTCACCA
289. GGTAAGCGTGGAAAGCCTGGGAAACGAGGAAAACCCGGTAAAAGGGGCAAGCCAGGGAAGAGAGGTAAACCA	799. GGTGGCAAGCATACCCCAGGAGGTAAGCACACTCCGGGAGGAAAACATACTCCAGGCGGTAAACACACACCA
290. GGTGGGAAGATGAGACCCGGCGGAAAAATGAGGCCAGGAGGTAAAATGCGTCCAGGGGGCAAAATGCGACCA	800. GGTGGTCACTGGAAACCTGGGGGACATTGGAAACCCGGCGGTCATTGGAAGCCAGGAGGGCATTGGAAACCA
291. GGTGGGAAGTTTAGGCCAGGAGGTAAATTTCGTCCAGGGGGCAAATTCAGACCCGGCGGAAAATTTCGACCA	801. GGTTATCATGGGAAGCCTGGATATCACGGCAAACCCGGTTACCACGGAAAGCCAGGATACCATGGAAAACCA
292. GGTGGTTCTAAAAGGCCAGGGGGATCAAAACGTCCAGGAGGTAGCAAGAGACCGGGCGGAAGTAAGCGACCA	802. GGTGGCAAGCATGTCCCAGGAGGAAAGCACGTACCGGGCGGAAAACACGTTCCAGGGGGTAAACATGTACCA
293. GGTCGCAAAGGAACCCCTGGAAGAAAAGGTACTCCCGGTAGAAAGGGCACGCCAGGACGTAAAGGCACACCA	803. GGTCATATGGGCTCTCCGGGACATATGGGTTCGCCAGGGCATATGGGAAGTCCAGGACACATGGGGTCACCA
294. GGTGGTTGGAAACGGCCAGGCGGATGGAAAAGACCGGGAGGGTGGAAGCGTCCTGGGGGCTGGAAACGACCA	804. GGTCACATGGGAACTCCGGGTCATATGGGTACTCCAGGGCATATGGGGACGCCAGGACATATGGGCACACCA
295. GGTAAAGGTTATCGCCCTGGCAAAGGCTATCGTCCTGGAAAAGGATACAGACCGGGTAAGGGATATCGACCA	805. GGTCACATGGGAGTACCCGGTCATATGGGCGTGCCAGGACATATGGGTGTTCCAGGGCATATGGGGGTACCA
296. GGTAAAGGCGTTAGACCGGGCAAAGGGGTTCGTCCAGGAAAAGGAGTACGGCCAGGGAAAGGTGTACGACCA	806. GGTCATGGATTTAGTCCTGGACACGGCTTTTCTCCCGGACATGGTTTTTCGCCAGGGCATGGGTTCTCACCA
297. GGTGGGATGATGAGGCCAGGAGGTATGATGAGACCCGGCGGAATGATGCGTCCAGGGGGCATGATGCGACCA	807. GGTTTTCATGGTACGCCCGGATTTCACGGCACTCCTGGCTTCCATGGACCCCTGGATTCCACGGTACACCA
298. GGTGGGATGTTTAGGCCAGGAGGTATGTTTCGTCCAGGGGGCATGTTCAGACCCGGCGGAATGTTTCGACCA	808. GGTCACTTTGGTGTACCGGGTCATTTTGGCGTTCCAGGACATTTCGGAGTCCCAGGGCACTTCGGCGTACCA
299. GGTGGTAGCATGAGGCCAGGAGGTTCTATGCGTCCAGGGGGATCAATGAGACCCGGCGGAAGTATGCGACCA	809. GGTTCACATGGCAGTCCAGGATCGCATGGATCTCCGGGAAGTCACGGTAGCCCTGGGTCTCATGGTTCACCA
300. GGTAGGATGGGAACGCCAGGACGTATGGGCACTCCGGGTCGCATGGGTACTCCAGGGAGAATGGGACACCA	810. GGTTCACATGGGACCCCTGGCTCTCATGGAACTCCTGGATCGCATGGTACGCCCGGAAGTCACGGTACACCA
301. GGTTGGGGCATGAGACCCGGCTGGGGGATGCGTCCAGGGTGGGGTATGAGGCCAGGATGGGGAATGCGACCA	811. GGTCATGGGTGGAGTCCCGGACACGGTTGGTCTCCTGGACATGGATGGTCGCCTGGGCATGGCTGGTCACCA
302. GGTATGTATGGTAGGCCAGGAATGTATGGGCGTCCAGGGATGTATGGAAGACCCGGCATGTACGGTCGACCA	812. GGTCATTATGGAAGTCCGGGACACTACGGCTCTCCAGGACATTACGGTTCGCCAGGGCACTATGGGTCACCA
303. GGTATGCGTGGAGTACCCGGAATGCGAGGAGTTCCTGGGATGAGAGGTGTTCCAGGCATGAGGGGCGTACCA	813. GGTGGCTCTCATGTGCCTGGAGGATCGCATGTCCCTGGGGGTTCACACGTTCCGGGCGGTAGTCATGTACCA
304. GGTTTTGGTTTCCGCCCAGGGTTCGGCTTTCGTCCAGGATTTGGCTTCAGACCGGGGTTTGGATTCCGACCA	814. GGTACACATGGCACTCCAGGAACTCACGGTACTCCGGGTACGCATGGAACCCCTGGGACTCATGGTACACCA
305. GGTGGGTTCAGTAGGCCAGGAGGTTTTAGCAGACCGGGCGGCTTTTCTCGTCCAGGGGGATTTAGTCGACCA	815. GGTCACTGGGGCACGCCAGGCCATTGGGGAACTCCGGGTCATTGGGGTACTCCTGGACATTGGGGGACACCA
306. GGTCGCTTTGGCACGCCTGGACGTTTTGGTACTCCCGGTAGATTTGGAACCCCAGGAAGATTCGGAACACCA	816. GGTGGCTATCATACTCCAGGCGGTTATCACACTCCGGGAGGATACCATACCCCAGGAGGTTACCACACACCA
307. GGTTGGTTTGGTAGGCCAGGGTGGTTTGGGCGTCCAGGATGGTTTGGAAGACCCGGCTGGTTCGGTCGACCA	817. GGTCACACTGGTGTTCCGGGTCATACAGGCGTGCCAGGACATACTGGAGTCCCAGGGCATACGGGGGTACCA
308. GGTTATGGTTTTAGACCGGGATATGGATTTCGCCCAGGGTATGGCTTTCGTCCAGGATACGGATTCCGACCA	818. GGTGGTTGGCATGTTCCTGGGGGCTGGCATGTACCGGGAGGGTGGCACGTTCCAGGCGGATGGCATGTACCA
309. GGTTTTCGTGGCGTCCCAGGATTTAGAGGAGTTCCAGGCTTTAGGGGTGTGCCCGGATTCAGAGGTGTACCA	819. GGTTACCATGGAGTGCCGGGGTATCACGGTGTTCCAGGATATCATGGCGTCCCAGGCTACCACGGAGTACCA
310. GGTTCTAGGGGCTCGCCAGGCAGTAGAGGAAGTCCCGGTAGTCGTGGGTCTCCTGGATCAAGAGGTTCACCA	820. GGTGGTGTCCATGTTCCTGGAGGAGTTCACGTACCGGGAGGGGTTCATGTGCCTGGCGGAGTACATGTACCA
311. GGTGGATCAACACGGCCTGGGGGAAGTACAAGACCCGGAGGTAGCACTCGTCCTGGCGGTTCTACTCGACCA	821. GGTATAATAGGGATTCCTGGAATTATTGGCATCCCAGGCATTATCGGAATACCGGGTATCATAGGTATACCA
312. GGTGGACGATGGAGCCCTGGCGGTAGATGGTCTCCCGGAGGCCGTTGGAGTCCTGGGGGTCGCTGGTCACCA	822. GGTGGCTTGATAATTCCTGGGGGTTTAATCATTCCGGGCGGTCTAATAATCCCTGGAGGACTTATAATACCA
313. GGTTACGGTAGATCGCCTGGGTATGGTCGTTCTCCTGGATATGGAAGAAGTCCGGGCTATGGCAGATCACCA	823. GGTGGCAAAATTATCCCAGGGGGTAAAATCATTCCGGGCGGTAAGATTATTCCAGGAGGAAAAATAATACCA
314. GGTAGAGTTGGATCGCCTGGCAGAGTAGGGAGCCCCGGACGAGTCGGTTCTCCAGGAAGAGTGGGTTCACCA	824. GGTGGAATTATGATACCCGGAGGTATTATGATTCCTGGCGGTATAATGATCCCTGGGGGCATCATGATACCA
315. GGTCGTGGTACAACCCCAGGCAGAGGTACTACGCCCGGTAGAGGAACTACTCCTGGAAGAGGGACTACACCA	825. GGTATATTCGGAATCCCTGGAATTTTCGGCATACCGGGTATCTTTGGCATTCCTGGGATATTTGGTATACCA
316. GGTAGATGGGGCACGCCAGGGAGGTGGGGTACTCCAGGACGTTGGGGAACTCCGGGTCGCTGGGGGACACCA	826. GGTATTGGCTCTATACCCGGAATAGGTTCAATTCCAGGAATCGGTAGTATCCCAGGCATAGGATCTATACCA
317. GGTCGCTATGGCACGCCTGGACGTTATGGTACTCCCGGTAGATACGGAACCCCAGGAAGATATGGAACACCA	827. GGTGGCATAATCACTCCCGGAGGTATAATAACCCCAGGCGGTATCATTACTCCAGGAGGAATTATCACACCA
318. GGTACAGTCGGAAGGCCAGGAACCGTTGGTAGACCCGGCACTGTGGGGCGTCCAGGGACAGTAGGTCGACCA	828. GGTTGGATCGGTATTCCTGGCTGGATAGGCATACCCGGATGGATAGGTATCCCTGGGTGGATTGGTATACCA
319. GGTTGGGGTTGGCGTCCAGGGTGGGGATGGAGACCCGGCTGGGGGTGGAGGCCAGGATGGGGCTGGCGACCA	829. GGTGGCATATACATTCCAGGAGGAATTTATATCCCAGGGGGTATCTATATTCCGGGCGGTATTTACATACCA
320. GGTTGGTACGGTAGACCCGGCTGGTATGGAAGGCCAGGATGGTATGGGCGTCCAGGGTGGTATGGTCGACCA	830. GGTGTAATTGGCATTCCGGGCGTCATCGGTATTCCAGGAGTGATAGGAATCCCTGGGGTTATAGGCATACCA
321. GGTGGTTGGGTAAGGCCAGGGGGATGGGTGAGACCGGGCGGCTGGGTTCGTCCTGGAGGGTGGGTCCGACCA	831. GGTATCTTAGGCTTGCCAGGAATACTAGGACTTCCAGGGATTCTCGGTCTACCCGGAATTTTGGGGTTACCA
322. GGTTATTATGGAAGGCCAGGATATTACGGACGCCCTGGGTACTACGGCAGACCGGGATACTATGGTCGACCA	832. GGTATTGGAAAGCTCCCAGGGATAGGCAAACTTCCAGGAATCGGTAAATTACCCGGAATAGGAAAATTACCA
323. GGTGGAAGATATGTGCCAGGAGGACGCTATGTCCCTGGGGGTCGTTATGTTCCGGGCGGTAGATACGTACCA	833. GGTGGGATTATGTTGCCAGGAGGAATAATGCTTCCAGGGGGTATCATGTTACCGGGCGGAATTATGTTACCA
324. GGTGGAAGAGTTGTTCCTGGCGGTCGTGTAGTACCCGGAGGTAGAGTAGTCCCAGGGGGCAGAGTGGTACCA	834. GGTGGCTTATTTATACCGGGCGGACTTTTTATCCCAGGGGGTCTATTCATTCCAGGAGGTTTGTTCATACCA
325. GGTAATAATGGGCAGCCAGGAAATAACGGACAGCCTGGAAACAATGGTCAACCGGGGAACAACGGCCAACCA	835. GGTGGCTTGTCAATCCCTGGAGGACTTTCTATTCCGGGCGGTCTTAGCATTCCTGGGGGTTTATCGATACCA
326. GGTGGAAACAATGAACCGGGCGGCAATAACGAGCCAGGAGGGAATAATGAGCCTGGGGGTAACAACGAACCA	836. GGTGGCTTGACGATTCCGGGCGGTTTAACAATTCCAGGAGGACTTACTATCCCAGGGGGTCTAACTATACCA
327. GGTGGCAACAATATACCGGGCGGAAATAATATCCCAGGGGGTAATAACATTCCAGGAGGAAACAACATACCA	837. GGTATTGGATGGTTGCCTGGGATAGGGTGGCTTCCTGGAATAGGTTGGTTACCCGGAATCGGCTGGTTACCA
328. GGTGGGAACAACTTGCCAGGAGGAAACAATCTTCCAGGGGGTAATAATCTACCGGGCGGAAATAACTTACCA	838. GGTGGCTTGTATATCCCAGGAGGACTCTATATTCCGGGCGGTCTTTACATTCCAGGGGGTTTATATATACCA
329. GGTAACAACGGAAAACCCGGTAATAATGGAAAGCCAGGAAATAACGGCAAGCCTGGGAACAATGGTAAACCA	839. GGTGTGATAGGTCTACCGGGAGTAATAGGATTGCCAGGCGTCATCGGACTTCCAGGGGTTATTGGGTTACCA
330. GGTAACAATGGAAGTCCAGGGAATAACGGCTCGCCCGGAAACAACGGTTCTCCTGGCAATAATGGGTCACCA	840. GGTATTAAGGGTAAGCCAGGAATTAAAGGAAAACCGGGGATAAAAGGCAAGCCTGGAATCAAAGGTAAACCA
331. GGTGGAAACAATACTCCGGGAGGAAATAACACTCCTGGCGGTAATAATACCCCTGGAGGCAACAACACACCA	841. GGTGGCAAAATGATTCCAGGGGGTAAAATGATCCCAGGAGGAAAGATGATACCGGGCGGAAAAATGATACCA
332. GGTAATGGTAACGTCCCTGGCAACGGCAATGTTCCAGGAAATGGAAACGTACCGGGGAATGGAACGTACCA	842. GGTTTTGGCAAGATCCCTGGATTTGGTAAAATTCCTGGGTTCGGAAAAATACCGGGCTTTGGAAAGATACCA
333. GGTGATAACGGACAGCCAGGGGATAATGGGCAGCCTGGAGACAATGGTCAACCCGGCGACAACGGTCAACCA	843. GGTGGCAGTAAGATACCGGGCGGAAGTAAAATCCCTGGGGGTTCAAAAATTCCTGGAGGCTCTAAAATACCA
334. GGTGGAGATAATGAACCGGGCGGAGACAATGAGCCAGGAGGGGATAACGAGCCTGGGGGTGACAACGAACCA	844. GGTGGCAAAACTATCCCAGGAGGAAAAACAATTCCAGGGGGTAAAACGATTCCGGGCGGTAAGACTATACCA
335. GGTGATGGCAATATCCCTGGGGACGGAAACATTCCTGGAGATGGTAATATACCCGGCGATGGAAATATACCA	845. GGTTGGGGGAAGATTCCTGGCTGGGGCAAAATACCGGGATGGGGTAAAATTCCAGGGTGGGGAAAGATACCA
336. GGTGGGAATGACTTGCCAGGGGGTAACGATCTTCCAGGAGGAAATGATCTACCGGGCGGAAACGACTTACCA	846. GGTGGCAAATACATACCGGGCGGAAAGTACATTCCAGGGGGTAAGTATATCCCAGGAGGAAAATATATACCA
337. GGTAACGACGGAAAGCCAGGAAATGATGGAAAACCCGGTAATGACGGCAAGCCTGGGAACGATGGTAAACCA	847. GGTGGCAAAGTAATCCCAGGAGGAAAAGTTATTCCGGGCGGTAAAGTCATTCCAGGGGGTAAGGTTATACCA
338. GGTGACAATGGCAGTCCCGGCGACAACGGTAGCCCTGGGGATAACGGAAGTCCTGGAGATAATGGTTCACCA	848. GGTATGGGTATGATACCCGGAATGGGCATGATCCCTGGGATGGGAATGATTCCTGGCATGGGGATGATACCA
339. GGTGGCAATGACACTCCAGGCGGTAATGATACTCCGGGAGGAAACGATACCCCAGGAGGTAACGACACACCA	849. GGTTTTGGGATGATACCCGGATTTGGAATGATCCCTGGATTCGGTATGATTCCTGGCTTTGGCATGATACCA
340. GGTGATAATGGAGTCCCTGGCGACAACGGTGTGCCCGGAGACAATGGGGTTCCAGGAGATAACGGCGTACCA	850. GGTATGATAGGGTCGCCTGGCATGATCGGTTCTCCGGGGATGATAGGAAGTCCTGGAATGATTGGGTCACCA
341. GGTCAAGGTAATCAACCGGGGCAAGGGAATCAGCCTGGACAGGGTAACCAACCCGGACAAGGAAATCAACCA	851. GGTGGCACAATGATTCCTGGAGGTACGATGATACCGGGCGGAACCATGATCCCAGGGGGTACTATGATACCA
342. GGTAACGAGGGGCAGCCAGGAAATGAAGGTCAACCCGGTAATGAGGGACAGCCTGGAAACGAAGGACAACCA	852. GGTTGGGGAATGATTCCTGGATGGGGGATGATACCGGGCTGGGGCATGATCCCTGGGTGGGGTATGATACCA
343. GGTAATGGTGGTCAGCCTGGAAATGGAGGACAGCCAGGGAACGGCGGACAACCGGGCAATGGGGGTCAACCA	853. GGTATGTATGGGATTCCGGGGATGTATGGTATTCCTGGAATGTACGGTATCCCTGGCATGTATGGAATACCA
344. GGTCATAATGGGCAGCCAGGACATAACGGACAGCCTGGACACAATGGTCAACCCGGTCACAACGGCCAACCA	854. GGTATAGGGATGGTGCCTGGAATCGGAATGGTTCCCGGTATTGGTATGGTCCCAGGAATTGGCATGGTACCA
345. GGTCAAAATGGAATCCCAGGGCAGAATGGCATACCCGGACAGAACGGTATTCCTGGACAAAACGGCATACCA	855. GGTTTTGGCTTTATTCCGGGGTTTGGTTTCATCCCTGGGTTCGGCTTCATTCCAGGATTTGGATTCATACCA
346. GGTAATCAAGGACTCCCAGGCAATCAAGGTCTACCCGGAAATCAGGGGCTTCCAGGAAACCAAGGGTTACCA	856. GGTTCATTTGGCATTCCGGGTAGTTTCGGTATTCCTGGATCGTTTGGAATCCCAGGCTCTTTTGGTATACCA
347. GGTAACGGAAAGCAGCCTGGGAATGGCAAACAACCGGGTAATGGTAAACAGCCAGGAAATGGAAAACAACCA	857. GGTACTTTCGGAATCCCTGGGACTTTTGGCATTCCTGGAACATTTGGAATACCCGGTACGTTTGGTATACCA
348. GGTGGAAACATGCAACCTGGGGGTAATATGCAACCGGGCGGAAATATGCAGCCAGGAGGGAATATGCAACCA	858. GGTTTTTGGGGAATACCGGGTTTCTGGGGTATCCCTGGCTTTTGGGGCATTCCAGGATTTTGGGGGATACCA
349. GGTAATTTTGGGCAGCCTGGAAACTTCGGACAACCCGGTAACTTTGGTCAGCCAGGGAATTTCGGCCAACCA	859. GGTTTCTACGGTATCCCTGGCTTTTACGGCATTCCTGGATTCTATGGAATACCCGGATTTTATGGGATACCA
350. GGTGGAAATCAGAGTCCTGGGGGTAATCAATCTCCCGGAGGCAACCAAAGCCCTGGCGGTAACCAGTCACCA	860. GGTGGCATATTTGTTCCAGGAGGAATTTTCGTACCGGGCGGAATCTTTGTCCCTGGGGGTATATTCGTACCA
351. GGTAATCAAGGGACCCCAGGCAATCAGGGTACGCCCGGAAACCAAGGAACTCCTGGAAATCAAGGTACACCA	861. GGTAGCATTGGAAGTCCTGGGTCTATCGGCTCTCCGGGATCAATAGGTTCGCCAGGCAGTATAGGGTCACCA
352. GGTGGAAACTGGCAACCTGGGGGTAATTGGCAACCGGGCGGAAATTGGCAGCCTGGAGGGAATTGGCAACCA	862. GGTGGAATTACATCTCCTGGCGGTATTACGAGCCCTGGGGGTATAACAAGTCCGGGAGGCATTACTTCACCA
353. GGTGGAAACTATCAGCCTGGGGGCAATTATCAACCCGGCGGTAACTACCAGCCAGGAGGGAATTACCAACCA	863. GGTGGCAGTTGGATCCCTGGGGGCTCATGGATACCCGGAGGATCGTGGATTCCAGGCGGTTCTTGGATACCA
354. GGTGTGAATGGTCAGCCTGGAGTTAATGGGCAACCCGGCGTAAATGGACAGCCAGGGGTAAACGGTCAACCA	864. GGTTATATCGGTTCGCCTGGCTACATAGGAAGTCCCGGATACATTGGATCTCCAGGGTATATTGGGTCACCA
355. GGTGGAGAGAATGAACCGGGCGGAGAAAATGAGCCAGGAGGGGAAAACGAGCCTGGGGGTGAGAACGAACCA	865. GGTGGCATTTCTGTACCGGGCGGAATAAGTGTCCCAGGGGGTATATCAGTTCCAGGAGGAATTAGCGTACCA
356. GGTGGAAACGGAGAGCCAGGAGGAAATGGTGAGCCTGGAGGTAATGGGGAACCGGGCGGCAATGGAGAACCA	866. GGTACAATCGGAACCCCAGGCACTATTGGTACTCCAGGGACCATAGGAACGCCCGGTACGATAGGTACACCA
357. GGTGGGAATCACGAGCCTGGAGGAAACCACGAACCCGGCGGTAACCATGAGCCAGGGGGCAATCATGAACCA	867. GGTTGGGGCACTATTCCAGGGTGGGGGACAATTCCGGGATGGGGTACGATTCCTGGCTGGGGAACTATACCA
358. GGTGAGGGTAACATTCCAGGAGAAGGAAATATACCGGGCGAAGGCAATATCCCAGGGGAAGGTAATATACCA	868. GGTTATATTGGGACCCCTGGCTATATCGGAACTCCTGGATACATAGGCACGCCCGGTTACATTGGTACACCA
359. GGTAACGAAGGTCTACCCGGAAATGAAGGACTTCCTGGGAATGAGGGCTTGCCTGGAAACGAGGGGTTACCA	869. GGTGTTGGCACTATTCCGGGCGTAGGTACTATCCCAGGGGTCGGTACAATTCCAGGAGTAGGAACTATACCA
360. GGTGGAGAGAATAAACCGGGCGGAGAAAATAAGCCAGGAGGGGAAAACAAGCCTGGGGGTGAGAACAAACCA	870. GGTGGCTGGTGGATACCGGGCGGATGGTGGATTCCTGGAGGGTGGTGGATCCCAGGGGGTTGGTGGATACCA
361. GGTGGTAACATGGAACCTGGGGGCAATATGGAACCCGGCGGAAATATGGAGCCAGGAGGGAATATGGAACCA	871. GGTGGCTGGTACATACCGGGCGGATGGTATATTCCAGGAGGGTGGTATATCCCAGGGGGTTGGTATATACCA
362. GGTGGAAACTTTGAGCCAGGAGGGAATTTTGAACCCGGCGGTAATTTCGAGCCTGGGGGCAACTTCGAACCA	872. GGTTGGATAGGCGTGCCCGGATGGATCGGTGTTCCTGGCTGGATAGGAGTCCCTGGGTGGATTGGTGTACCA
363. GGTGGAAACGAAAGTCCGGGAGGCAATGAGAGCCCTGGCGGTAATGAATCTCCTGGGGGTAACGAGTCACCA	873. GGTTATTACGGCATTCCTGGCTACTACGGAATACCGGGGTACTATGGAATCCCTGGATATTATGGGATACCA
364. GGTGGCAACGAAACTCCCGGAGGAAATGAGACTCCAGGCGGTAATGAAACCCCAGGAGGTAACGAGACACCA	874. GGTGGAATATACGTTCCGGGCGGTATTTATGTGCCAGGGGGTATCTATGTCCCAGGAGGAATTTACGTACCA
365. GGTGGGTGGAATGAACCGGGCGGCTGGAACGAGCCAGGAGGTTGGAATGAGCCTGGGGGATGGAATGAACCA	875. GGTGTAGGAATAGTTCCGGGCGTAGGTATAGTGCCAGGGGTTGGAATTGTCCCAGGAGTCGGTATCGTACCA
366. GGTTACAATGGTGAGCCTGGGTATAACGGCGAACCGGGATATAATGGAGAGCCAGGATACAACGGAGAACCA	876. GGTTTGTTGGGATTGCCAGGCTTACTCGGACTTCCAGGATTATTAGGTCTACCGGGGTTACTAGGGTTACCA
367. GGTGAAGGTAATGTTCCAGGAGAAGGCAATGTCCCTGGGGAAGGAAATGTACCGGGCGAGGGTAACGTACCA	877. GGTTTGAAAGGTCTTCCAGGGTTAAAAGGCTTGCCAGGACTTAAAGGATTACCGGGATTGAAGGGGTTACCA
368. GGTGGTGGAAATATACCGGGAGGGGGTAACATTCCTGGCGGAGGTAATATCCCTGGAGGAGGCAATATACCA	878. GGTGGGATGTTACTTCCTGGCGGTATGCTACTACCCGGAGGAATGCTTCTTCCAGGGGGCATGTTGTTACCA
369. GGTGGTGGAAATCTTCCTGGAGGCGGTAACTTGCCAGGGGGAGGTAATCTACCCGGAGGAGGGAATTTACCA	879. GGTTTTCTCGGACTTCCAGGATTCTTAGGTCTACCGGGATTTTTGGGGTTGCCTGGCTTTTTAGGGTTACCA
370. GGTGGTAATGGGAAGCCTGGAGGAAATGGAAAGCCAGGGGGTAACGGAAAACCGGGCGGCAATGGTAAACCA	880. GGTCTTGGATTATCTCCGGGCTTAGGTTTATCCCCAGGACTCGGTTTGTCGCCAGGGTTAGGATTGTCACCA
371. GGTGGTAACGGTAGTCCGGGAGGAAATGGAAGCCCAGGCGGCAATGGTTCTCCAGGGGGTAATGGCTCACCA	881. GGTACGCTCGGATTGCCTGGAACTCTAGGTTTACCGGGTACACTTGGTCTTCCTGGGACATTAGGGTTACCA
372. GGTGGTAACGGTACTCCTGGCGGCAATGGAACCCCTGGAGGAAATGGGACTCCGGGAGGTAATGGTACACCA	882. GGTGGTCTCTGGCTTCCTGGGGGACTTTGGTTGCCTGGAGGTTTATGGTTACCGGGCGGATTGTGGTTACCA
373. GGTGGTGGAAATGTACCGGGAGGGGGTAATGTTCCTGGCGGAGGCAATGTCCCTGGAGGAGGTAACGTACCA	883. GGTTTGTATGGACTCCCCGGTCTTTATGGGCTTCCTGGATTATACGGATTGCCAGGGTTATATGGTTTACCA
374. GGTGGGCATAATATCCCTGGCGGTCACAATATACCCGGAGGTCATAACATTCCTGGAGGACACAACATACCA	884. GGTTTGTTAGGTGTCCCTGGATTACTAGGCGTGCCCGGTCTTCTCGGAGTTCCTGGCTTATTGGGTGTACCA
375. GGTAATCATGGACTACCCGGAAATCACGGTCTTCCAGGGAACCACGGCTTGCCAGGAAACCATGGGTTACCA	885. GGTGGGAAAAAACTTCCAGGAGGTAAGAAATTGCCAGGGGGTAAAAAGCTACCGGGCGGAAAGAAGTTACCA
376. GGTGGACACAATAAGCCAGGAGGGCATAACAAGCCTGGGGGACATAATAAACCCGGCGGTCACAACAAACCA	886. GGTGGGATGAAACTTCCAGGGGGCATGAAGCTACCCGGAGGAATGAAATTGCCAGGCGGTATGAAATTACCA

377.	GGTGGCAACCATAGCCCTGGGGGTAATCACAGTCCCGGAGGAAATCATTCTCCAGGCGGTAACCACTCACCA	887.	GGTTTTGGCAAACTCCCTGGGTTCGGAAAATTACCGGGCTTTGGGAAACTTCCTGGATTTGGAAAGTTACCA
378.	GGTCACAATGGAACTCCCGGTCATAACGGAACCCCAGGACATAATGGTACGCCAGGGCACAACGGCACACCA	888.	GGTAAGAGTGGTCTCCCAGGAAAAAGCGGACTTCCAGGCAAATCAGGGCTACCGGGTAAATCTGGCTTACCA
379.	GGTGGAAACCATGTGCCAGGAGGAAATCACGTCCCAGGGGGTAATCATGTTCCGGGCGGTAACCACGTACCA	889.	GGTGGCAAAACTCTCCCTGGGGGTAAAACATTACCGGGCGGAAAGACACTTCCAGGAGGAAAAACGTTACCA
380.	GGTAATATCGGAATACCCGGTAACATAGGAATCCCAGGGAACATTGGCATTCCAGGAAATATAGGCATACCA	890.	GGTGGTTGGAAACTTCCTGGGGGCTGGAAATTGCCTGGCGGATGGAAGCTACCGGGAGGGTGGAAATTACCA
381.	GGTGGCTTAAATATCCCTGGGGGTTTGAACATTCCTGGAGGTCTTAATATACCGGGCGGACTAAATATACCA	891.	GGTAAGGGGTATCTTCCAGGGAAAGGTTATCTCCCAGGAAAAGGATATTTACCGGGCAAAGGCTATTTACCA
382.	GGTGGCAAAAATATACCGGGCGGAAAGAATATCCCTGGGGGTAAGAACATTCCTGGAGGTAAAAACATACCA	892.	GGTAAAGGTGTTCTCCCAGGAAAAGGAGTGCTACCGGGCAAAGGCGTACTTCCAGGGAAAGGGGTCTTACCA
383.	GGTGGCAATATGATACCGGGCGGAAACATGATTCCAGGAGGAAATATGATCCCTGGGGGTAATATGATACCA	893.	GGTGGCATGATGCTTCCTGGGGGTATGATGTTGCCTGGAGGGATGATGTTACCGGGCGGAATGATGTTACCA
384.	GGTGGCAATTTTATCCCAGGAGGAAACTTTATTCCGGGCGGTAATTTCATTCCAGGGGGTAACTTCATACCA	894.	GGTTTTATGGGACTTCCAGGATTCATGGGTCTACCCGGATTTATGGGGTTGCCTGGCTTTATGGGCTTACCA
385.	GGTAGTAACGGTATTCCTGGATCTAATGGCATACCCGGAAGCAATGGAATCCCAGGCTCAAATGGTATACCA	895.	GGTATGCTCGGTTCGCCAGGCATGTTGGGTAGTCCGGGGATGCTAGGATCTCCTGGAATGTTAGGGTCACCA
386.	GGTAATATCGGTACGCCCGGAAACATAGGAACCCCTGGCAACATTGGGACTCCTGGAAATATAGGCACACCA	896.	GGTGGGACAATGTTGCCAGGGGGTACTATGTTACCGGGCGGAACCATGCTTCCTGGAGGTACGATGTTACCA
387.	GGTGGAAATTGGATTCCTGGGGGCAATTGGATCCCTGGCGGTAACTGGATACCCGGAGGTAATTGGATACCA	897.	GGTGGTTGGATGCTTCCTGGGGGCTGGATGTTACCGGGAGGGTGGATGTTGCCTGGCGGATGGATGTTACCA
388.	GGTAACTACGGTATCCCTGGAAATTACGGCATTCCTGGCAATTATGGAATACCCGGAAACTATGGGATACCA	898.	GGTATGTATGGACTTCCTGGCATGTATGGTTTACCCGGAATGTACGGTCTTCCAGGGATGTATGGGTTACCA
389.	GGTGGAAATATAGTCCCTGGGGGTAATATCGTTCCGGGCGGTAACATTGTGCCAGGAGGAAACATAGTACCA	899.	GGTGGGATGCTCGTTCCTGGCGGTATGCTTGTACCCGGAGGAATGTTAGTTCCAGGGGGCATGTTGGTACCA
390.	GGTGGGAATTTATTGCCAGGAGGAAACTTACTACCGGGCGGAAATCTGCTTCCAGGGGGTAATTTGTTACCA	900.	GGTTTCTTCGGTCTCCCTGGCTTTTTTGGACTTCCAGGATTTTTCGGATTGCCGGGGTTCTTTGGGTTACCA
391.	GGTGGAAAAAATCTTCCTGGCGGTAAAAACTTGCCTGGGGGAAAGAATCTACCCGGAGGTAAGAACTTACCA	901.	GGTCTTGGATTTTCGCCCGGATTAGGCTTTTCCCCAGGACTCGGTTTTTCTCCAGGGTTAGGATTCTCACCA
392.	GGTGGGAATATGTTGCCAGGGGGTAATATGCTACCGGGCGGAAACATGCTTCCAGGAGGAAATATGTTACCA	902.	GGTGGGACATTTCTTCCAGGGGGTACGTTTTTGCCAGGAGGTACTTTTCTACCGGGCGGAACATTCTTACCA
393.	GGTTTTAACGGCTTGCCAGGATTCAACGGACTTCCAGGGTTCAATGGTCTACCGGGATTTAATGGGTTACCA	903.	GGTTGGTTCGGACTTCCAGGATGGTTTGGATTGCCAGGGTGGTTTGGTCTCCCCGGCTGGTTTGGGTTACCA
394.	GGTGGGAACTCATTGCCAGGGGGTAATAGTCTTCCAGGAGGAAATTCTCTACCGGGCGGAAACAGCTTACCA	904.	GGTTATTTTGGACTCCCCGGCTATTTCGGTCTTCCAGGGTACTTCGGATTGCCAGGATACTTTGGGTTACCA
395.	GGTCTAAATGGTACGCCTGGACTTAATGGCACTCCCGGCTTGAATGGAACCCCAGGATTAAACGGTACACCA	905.	GGTTTTGGTGTATTGCCAGGATTCGGCGTACTACCGGGATTTGGAGTTCTTCCAGGCTTTGGGGTCTTACCA
396.	GGTTGGAATGGTCTACCGGGATGGAACGGACTTCCTGGGTGGAATGGATTGCCTGGCTGGAATGGGTTACCA	906.	GGTTTATCAGGCAGCCCAGGGTTGAGCGGATCTCCAGGATTAAGTGGTAGTCCCGGTCTTTCTGGCTCACCA
397.	GGTGGGAATTACTTGCCAGGGGGTAACTATCTTCCAGGAGGAAATTATCTACCGGGCGGAAACTACTTACCA	907.	GGTTCTCTAGGCACGCCCGGAAGTCTTGGGACCCCTGGATCACTCGGTACTCCTGGCTCATTAGGTACACCA
398.	GGTCTTAATGGAGTGCCTGGACTAAATGGTGTTCCGGGTTTGAATGGCGTCCCAGGATTAAACGGTGTACCA	908.	GGTTGGCTCGGTTCGCCTGGATGGCTTGGTAGTCCTGGGTGGTTAGGATCTCCGGGCTGGCTAGGGTCACCA
399.	GGTAAAAACGGAAAGCCAGGGAAGAATGGAAACCCGGCAAGAACGGTAAGCCTGGAAAAAATGGTAAACCA	909.	GGTTATCTTGGATCGCCTGGCTATCTCGGTTCTCCTGGGTATCTAGGAAGTCCCGGATATTTAGGGTCACCA
400.	GGTGGGAATATGAAGCCAGGGGGTAATATGAAACCGGGCGGCAACATGAAACCTGGAGGAAATATGAAACCA	910.	GGTAGTGTCGGACTCCCTGGAAGCGTAGGATTGCCTGGCTCAGTTGGTCTACCGGGTTCTGTGGGGTTACCA
401.	GGTAACTTTGGTAAGCCTGGGAATTTTGGAAAACCCGGCAATTTCGGAAAGCCAGGAAACTTCGGTAAACCA	911.	GGTGGGACAACATTGCCAGGGGGTACTACTCTACCGGGCGGAACTACACTTCCTGGAGGTACGACTTTACCA
402.	GGTGGAAACAAATCTCCTGGGGGTAATAAGAGTCCCGGAGGCAATAAAAGCCCTGGCGGTAACAAGTCACCA	912.	GGTGGTTGGACCTTGCCTGGCGGATGGACTCTACCGGGAGGGTGGACACTTCCTGGGGGCTGGACGTTACCA
403.	GGTGGCAACAAGACTCCAGGCGGTAACAAAACTCCGGGAGGTAATAAACCCCAGGAGGAAATAAGACACCA	913.	GGTACTTACGGTCTTCCAGGGACTTATGGGCTACCCGGAACCTATGGATTGCCAGGAACGTATGGTTTACCA
404.	GGTGGGTGGAATAAACCGGGCGGCTGGAATAAGCCAGGAGGTTGGAACAAGCCTGGGGGATGGAATAAACCA	914.	GGTCTAGTGGGCACGCCTGGCTTAGTAGGAACCCCTGGATTAGTTGGTACTCCGGGACTTGTCGGTACACCA
405.	GGTAACTATGGTAAGCCAGGAAATTACGGAAAACCCGGCAATTATGGAAAGCCTGGGAACTACGGTAAACCA	915.	GGTTGGTGGGGACTACCCGGATGGTGGGGCTTGCCAGGGTGGTGGGGTCTTCCAGGCTGGTGGGGGTTACCA
406.	GGTGGAAACAAGGTCCCTGGGGGTAACAAAGTTCCGGGCGGTAATAAAGTGCCAGGAGGAAATAAGGTACCA	916.	GGTTGGTACGGATTGCCAGGGTGGTATGGTCTCCCCGGCTGGTATGGACTTCCAGGATGGTATGGGTTACCA
407.	GGTAACATGGGCTCGCCAGGCAATATGGGTTCTCCGGGTAATATGGGAAGTCCAGGAAATATGGGGTCACCA	917.	GGTGTAGGTTGGTTACCGGGCGTAGGATGGCTTCCAGGAGTAGGGTGGTTGCCTGGGGTCGGCTGGTTACCA
408.	GGTGGGATGAACACTCCAGGGGGCATGAATACACCCGGAGGAATGAATACTCCTGGCGGTATGAATACACCA	918.	GGTTACGGCTACCTTCCAGGGTATGGTTATCTCCCCGGATATGGGTACTTGCCAGGCTATGGATACTTACCA
409.	GGTGGAAATATGGTTCCGGGCGGTAATATGGTGCCAGGGGGTAACATGGTCCCAGGAGGCAATATGGTACCA	919.	GGTTATCTAGGAGTCCCCGGATATTTGGGCGTTCCAGGCTACTTAGGTGTGCCAGGGTATCTCGGTGTACCA
410.	GGTGGAAACTTTAGTCCCGGAGGCAATTTTTCTCCTGGGGGTAACTTCAGCCCTGGCGGTAATTTCTCACCA	920.	GGTGGTGTACTTGTTCCTGGAGGAGTTCTCGTACCGGGAGGGGTTTTAGTGCCAGGCGGAGTCTTAGTACCA
411.	GGTGGATTTAATACTCCCGGAGGATTCAATACCCCTGGCGGTTTTAACACTCCTGGAGGCTTCAACACACCA	921.	GGTAAAAAGGGCAAGCCAGGGAAGAAGGGTAAACCCGGTAAGAAAGGAAAGCCTGGAAAAAAAGGTAAACCA
412.	GGTTTCAACGGTGTGCCAGGCTTTAACGGCGTCCCAGGATTTAATGGAGTTCCGGGGTTCAATGGTGTACCA	922.	GGTATGAAAGGCAAGCCTGGAATGAAAGGAAAACCGGGCATGAAGGGTAAGCCAGGGATGAAAGGTAAACCA
413.	GGTAATGGATCTTCTCCGGGGAATGGGTCTAGTCCAGGAAATGGTTCATCGCCAGGCAACGGCTCTTCACCA	923.	GGTAAATTCGGCAAGCCTGGAAAATTTGGTAAACCGGGTAAGTTTGGGAAGCCAGGAAAGTTCGGAAAACCA
414.	GGTAATAGCGGTACTCCAGGCAATTCAGGGACGCCCGGAAACAGTGGAACCCCAGGAAATTCTGGCACACCA	924.	GGTAAAAAGGGTAGTCCGGGCAAAAAAGGATCTCCTGGGAAGAAAGGTTCGCCTGGAAAGAAGGGGTCACCA
415.	GGTGGAAATTGGTCTCCTGGCGGTAACTGGAGTCCCGGAGGCAATTGGAGCCCTGGGGGTAATTGGTCACCA	925.	GGTACTAAGGGTAAGCCAGGAACTAAAGGAAAACCGGGGACAAAAGGCAAGCCTGGAACGAAAGGTAAACCA
416.	GGTTATGGAAATAGTCCGGGGTATGGTAATTCGCCAGGATATGGGAACTCTCCAGGCTACGGCAATTCACCA	926.	GGTGGGTGGAAAAAACCGGGCGGCTGGAAGAAGCCAGGAGGTTGGAAAAAGCCTGGGGGATGGAAAAAACCA
417.	GGTGTAGGAAATAGTCCCGGAGTTGGAAACAGCCCTGGCGTAGGCAATTCTCCTGGGGTCGGTAATTCACCA	927.	GGTAAATACGGCAAGCCAGGGAAATATGGAAAACCCGGAAAGTACGGTAAGCCTGGGAAGTATGGTAAACCA
418.	GGTAATACAGGGACCCCAGGAAATACTGGTACTCCGGGTAACACTGGAACGCCAGGCAATACGGGCACACCA	928.	GGTGGCAAAAAGGTCCCTGGGGGTAAGAAAGTACCGGGCGGAAAGAAGGTTCCAGGAGGTAAAAAAGTACCA
419.	GGTGGCTGGAATACCCCAGGAGGATGGAACACTCCGGGAGGGTGGAATACTCCAGGCGGTTGGAATACACCA	929.	GGTGGAATGATGAAGCCAGGAGGGATGATGAAACCCGGCGGTATGATGAAACCTGGGGGCATGATGAAACCA
420.	GGTTATGGAAACACGCCAGGATACGGCAATACTCCGGGCTATGGTAATACCCCTGGATATGGAATACACCA	930.	GGTATGTTTGGAAAGCCAGGAATGTTCGGAAAACCCGGCATGTTTGGTAAGCCTGGGATGTTTGGAAACCA
421.	GGTAATGTTGGGACGCCCGGAAATGTCGGTACTCCTGGCAATGTAGGAACCCCTGGAAACGTAGGCACACCA	931.	GGTGGGATGAGTAAGCCTGGAGGAATGTCAAAGCCAGGGGGTATGTCTAAACCGGGCGGCATGTCGAAACCA
422.	GGTTGGGGGAATGTACCGGGATGGGGAAATGTGCCTGGCTGGGGTAATGTTCCAGGGTGGGGCAATGTACCA	932.	GGTATGAAAGGAACCCCTGGCATGAAGGGTACGCCCGGAATGAAAGGTACTCCTGGGATGAAAGGCACACCA
423.	GGTTACAATGGTGTGCCCGGATATAATGGAGTTCCTGGGTATAACGGCGTCCCTGGCTACAACGGTGTACCA	933.	GGTGGTATGTGGAAGCCAGGAGGGATGTGGAAACCCGGCGGAATGTGGAAACCTGGGGGCATGTGGAAACCA
424.	GGTGGCAATGTCGTCCCAGGAGGAAACGTAGTTCCAGGGGGTAATGTTGTACCGGGCGGAAATGTGGTACCA	934.	GGTTATATGGGTAAGCCTGGATATATGGGAAAACCCGGCTATATGGGCAAGCCAGGATACATGGGGAAACCA
425.	GGTGATGACGGACAGCCAGGGGACGATGGGCAACCCGGCGATGATGGTCAGCCTGGAGACGACGGTCAACCA	935.	GGTAAGGGGATGGTGCCAGGAAAAGGCATGGTCCCAGGGAAAGGTATGGTTCCGGGTAAAGGAATGGTACCA
426.	GGTGATGGTGACGAGCCTGGAGACGGCGATGAACCGGGCGACGGAGATGAGCCAGGGGATGGGGACGAACCA	936.	GGTTTCTTCGGAAAGCCAGGATTCTTTGGTAAACCGGGATTTTTCGGCAAGCCTGGGTTTTTTGGAAAACCA
427.	GGTGATGATGGCATTCCTGGGGACGATGGAATACCCGGCGATGACGGAATCCCTGGAGACGACGGTATACCA	937.	GGTGGAAAATTCTCTCCTGGCGGTAAGTTTAGCCCTGGGGGTAAATTCAGTCCCGGAGGCAAATTTTCACCA
428.	GGTGACGGCGACTTGCCAGGAGATGGGGACCTTCCAGGGGATGGTGATCTACCGGGCGATGGAGACTTACCA	938.	GGTAAATTCGGTACTCCAGGAAAGTTTGGGACGCCCGGAAAATTTGGAACCCCTGGCAAGTTCGGCACACCA
429.	GGTGGGGACGACAAGCCTGGAGGAGATGATAAACCGGGCGGAGACGATAAGCCAGGGGGTGATGACAAACCA	939.	GGTTTTTGGGGTAAACCGGGATTTTGGGGCAAGCCTGGATTCTGGGGAAAGCCAGGGTTTTGGGGGAAACCA
430.	GGTGATGACGGAAGTCCTGGAGATGATGGTAGCCCAGGGGACGACGGTTCGCCGGGAGACGATGGCTCACCA	940.	GGTTATTTCGGCAAGCCAGGATATTTTGGTAAGCCTGGGTACTTCGGAAAACCGGGATACTTTGGAAAACCA
431.	GGTGATGACGGCACTCCTGGAGACGATGGAACCCCAGGGGACGACGGTACGCCCGGAGATGATGGCACACCA	941.	GGTGGAAAATTTGTTCCGGGCGGTAAGTTCGTCCCAGGAGGAAAGTTTGTGCCAGGGGGTAAATTCGTACCA
432.	GGTGGGGATGATGTACCGGGCGGAGACGACGTGCCAGGGGGTGACGATGTTCCTGGAGGAGATGACGTACCA	942.	GGTAGTAAGGGCAGTCCAGGGTCTAAAGGATCGCCAGGAAGCAAAGGCTCTCCCGGTTCAAAAGGTTCACCA
433.	GGTGATCAAGGTCAACCCGGCGATCAAGGACAGCCAGGAGACCAAGGGCAGCCTGGGGATCAGGGTCAACCA	943.	GGTGGAACTAAAAGCCCTGGGGGTACAAAATCTCCTGGCGGTACTAAGAGTCCCGGAGGCACGAAATCACCA
434.	GGTGGGCAAGATGAACCCGGCGGTCAGGATGAGCCTGGAGGTCAAGACGAGCCAGGGGGACAGGACGAACCA	944.	GGTAAATGGGGAAGTCCAGGAAAATGGGGCTCTCCGGGTAAGTGGGGTTCGCCTGGCAAATGGGGGTCACCA
435.	GGTGGTGATGGACAGCCTGGAGGGGATGGTCAACCCGGCGGAGATGGGCAGCCAGGAGGAGACGGTCAACCA	945.	GGTTATGGTAAATCGCCTGGATATGGAAAAAGTCCTGGCTACGGCAAATCTCCGGGGTATGGGAAGTCACCA
436.	GGTGGGCATGACCAGCCTGGAGGACACGACCAACCCGGCGGTCACGATCAGCCAGGGGGACATGATCAACCA	946.	GGTAAAGTTGGATCTCCAGGCAAAGTGGGTAGTCCGGGGAAGGTCGGTTCGCCAGGAAAAGTAGGGTCACCA
437.	GGTGGAATAGACCAGCCTGGGGGCATTGACCAACCCGGCGGTATAGATCAGCCAGGAGGGATTGATCAACCA	947.	GGTGGCAAAACTACTCCAGGCGGTAAAACGACTCCGGGAGGTAAGACAACCCCAGGAGGAAAAACAACACCA
438.	GGTCAAGGTGATTTGCCAGGCCAAGGGGATCTACCGGGTCAGGGCGATCTTCCTGGACAAGGAGACTTACCA	948.	GGTGGTACGTGGAAACCTGGGGGCACTTGGAAACCCGGCGGAACATGGAAGCCAGGAGGGACTTGGAAACCA
439.	GGTAAAGACGGACAGCCTGGAAAAGATGGTCAACCCGGTAAGGATGGGCAGCCAGGGAAGGACGGCCAACCA	949.	GGTTACAAGGGAACTCCGGGGTATAAGGGCACGCCAGGCTACAAAGGTACTCCAGGATATAAAGGAACACCA
440.	GGTGGGGATATGCAGCCAGGGGGTGATATGCAACCGGGCGGAGACATGCAACCTGGAGGAGATATGCAACCA	950.	GGTGGTGTAACGAAGCCTGGAGGAGTAACAAAACCGGGCGGAGTCACTAAGCCAGGGGGAGTTACCAAACCA
441.	GGTTTTGACGGACAGCCTGGATTCGATGGGCAACCGGGATTTGATGGTCAGCCAGGGTTCGACGGCCAACCA	951.	GGTTGGTGGGGAAAGCCTGGATGGTGGGGCAAACCCGGCTGGTGGGGTAAGCCAGGGTGGTGGGGGAAACCA
442.	GGTAGTGATGGGCAGCCTGGATCGGATGGTCAACCGGGTTCAGATGGACAGCCAGGAAGCGACGGCCAACCA	952.	GGTGGGTGGTATAAGCCAGGGGGATGGTACAAGCCTGGAGGTTGGTATAAACCGGGCGGCTGGTATAAACCA
443.	GGTGACCAAGGTACACCCGGAGATCAAGGAACTCCTGGGGATCAGGGCACTCCAGGCGATCAAGGGACACCA	953.	GGTGGTGTTTGGAAACCGGGAGGAGTATGGAAGCCTGGCGGAGTGTGGAAACCTGGAGGGGTCTGGAAACCA
444.	GGTGATTGGGGCCAACCCGGCGATTGGGGTCAGCCTGGAGACTGGGGACAGCCAGGAGATTGGGGGCAACCA	954.	GGTTATTACGGCAAGCCTGGGTACTATGGAAAACCGGGATATTATGGTAAGCCAGGATACTACGGAAAACCA
445.	GGTGATTATGGGCAGCCAGGGGATTACGGACAGCCTGGAGACTATGGTCAACCCGGCGACTACGGTCAACCA	955.	GGTTATGTAGGAAAGCCAGGGTATGTGGGCAAACCGGGATATGTTGGTAAGCCTGGATACGTCGGAAAACCA
446.	GGTGTTGATGGGCAGCCTGGAGTCGATGGACAACCCGGCGTAGACGGTCAGCCAGGGGTAGATGGTCAACCA	956.	GGTGGCAAAGTCGTCCCTGGAGGTAAGGTTGTACCGGGCGGAAAAGTAGTTCCTGGGGGTAAAGTGGTACCA
447.	GGTGGAGAGGATGAACCGGGCGGAGAAGATGAGCCAGGAGGGGAAGACGAGCCTGGGGGTGAGGACGAACCA	957.	GGTATGATGGGAAGTCCAGGGATGATGGGTTCTCCCGGAATGATGGGCTCGCCAGGCATGATGGGGTCACCA
448.	GGTGATGGAGGAGAGCCTGGAGACGGCGGAGAACCGGGAGATGGGGGTGAGCCAGGGGATGGTGGTGAACCA	958.	GGTATGATGGGAACGCCAGGGATGATGGGTACTCCCGGCATGATGGGGACTCCTGGAATGATGGGCACACCA
449.	GGTGACCACGGAGAGCCTGGAGACCATGGGGAACCCGGCGATCACGGTGAGCCAGGAGATCATGGAGAACCA	959.	GGTATGATGGGAGTTCCAGGGATGATGGGCGTACCCGGAATGATGGGTGTTCCTGGCATGATGGGGGTACCA
450.	GGTGAAGACGGAATCCCTGGGGAAGATGGCATTCCTGGAGAGGATGGAATACCCGGCGAGGACGGTATACCA	960.	GGTATGTTCGGTTCTCCGGGGATGTTTGGGTCGCCAGGCATGTTTGGAAGTCCAGGAATGTTTGGTTCACCA
451.	GGTTTAGACGGAGAGCCTGGACTCGATGGGGAACCGGGTCTTGATGGTGAGCCAGGATTAGATGGAGAACCA	961.	GGTATGTTCGGTACTCCAGGGATGTTTGGTACGCCGGGCATGTTTGGAACCCCAGGAATGTTTGGCACACCA
452.	GGTGAGGGAGATAAACCCGGAGAAGGAGACAAGCCTGGAGAGGGTGATAAGCCAGGGGAAGGCGATAAACCA	962.	GGTTTTGGAATGGTCCCTGGATTTGGCATGGTGCCCGGCTTCGGTATGGTTCCTGGGTTTGGGATGGTACCA
453.	GGTGATGGCATGGAACCTGGGGACGGAATGGAACCCGGCGATGGTATGGAGCCAGGAGATGGGATGGAACCA	963.	GGTAGTATGGGTTCGCCTGGATCAATGGGCAGTCCTGGCTCGATGGGATCTCCCGGAAGCATGGGGTCACCA
454.	GGTTTTGATGGTGAGCCAGGATTTGACGGAGAACCCGGTTTCGACGGCGAGCCTGGATTCGATGGGGAACCA	964.	GGTAGCATGGGAACTCCTGGATCTATGGGCACTCCGGGTTCAATGGGTACGCCTGGCTCGATGGGGACACCA
455.	GGTGATGAGGGTTCTCCAGGAGACGAAGGTAGTCCCGGAGATGAAGGAAGCCCAGGCGACGAGGGGTCACCA	965.	GGTGGCATGTGGTCTCCCGGCGGAATGTGGAGTCCAGGAGGGATGTGGAGCCCTGGGGGTATGTGGTCACCA
456.	GGTGGTGATACAGAGCCTGGAGGGGATACCGAACCCGGCGGAGATACTGAGCCAGGAGGAGACACGGAACCA	966.	GGTATGTATGGAAGTCCGGGCATGTATGGGTCTCCAGGAATGTATGGTTCGCCAGGGATGTACGGTTCACCA
457.	GGTGGATGGGATGAGCCTGGAGGGTGGGATGAACCGGGCGGCTGGGACGAGCCAGGAGGTTGGGATGAACCA	967.	GGTGGGATGTCAGTTCCTGGCGGTATGTCTGTACCCGGAGGAATGAGTGTTCCAGGGGGCATGAGCGTACCA
458.	GGTGGAGACTATGAACCGGGCGGAGATTACGAGCCAGGAGGGGATTATGAGCCTGGGGGTGACTACGAACCA	968.	GGTACTATGGGTACGCCAGGGACCATGGGAACTCCAGGAACGATGGGCACTCCCGGTACAATGGGGACACCA
459.	GGTGGAGACGTAGAACCGGGCGGAGATGTTGAGCCAGGAGGGGATGTAGAGCCTGGGGGTGATGTCGAACCA	969.	GGTTGGATGGGAACTCCAGGGTGGATGGGCACGCCAGGCTGGATGGGTACTCCCGGATGGATGGGGACACCA
460.	GGTGGTGATGGCATTCCTGGAGGCGATGGAATACCCGGAGGAGACGGTATCCCAGGGGGAGATGGTATACCA	970.	GGTATGTACGGAACCCCTGGGATGTATGGAACGCCGGGCATGTATGGTACTCCAGGAATGTATGGCACACCA
461.	GGTGGTGACGGCTTGCCTGGAGGGGATGGTCTTCCTGGCGGTGATGGACTACCCGGAGGAGATGGGTTACCA	971.	GGTATGGTGGGTACTCCAGGCATGGTCGGTACGCCAGGAATGGTTGGAACTCCGGGGATGGTAGGGACACCA
462.	GGTGGTGATGGAAAGCCAGGAGGAGACGGCAAACCGGGCGGAGATGGTAAGCCTGGAGGGGATGGGAAACCA	972.	GGTTGGATGGGGGTTCCAGGGTGGATGGGCGTACCCGGCTGGATGGGAGTGCCAGGATGGATGGGTGTACCA
463.	GGTGGAGACGGTAGCCCAGGAGGAGATGGCAGTCCCGGAGGCGATGGAAGTCCAGGGGGTGATGGTTCACCA	973.	GGTGGGATGTACGTTCCGGGCGGCATGTATGTTCCAGGGGGTATGTATGTGCCAGGAGGAATGTATGTACCA
464.	GGTGGAGATGGAACCCCAGGAGGAGACGGCACTCCCGGAGGGGATGGTACTCCAGGCGGTGATGGCACACCA	974.	GGTGGAGTGATGGTTCCTGGGGGTGTCATGGTCCCTGGCGGTGTAATGGTGCCCGGAGGAGTTATGGTACCA
465.	GGTGGTGGAGACGTACCGGGAGGAGGTGATGTTCCTGGAGGCGGAGATGTGCCAGGGGGTGGGGATGTACCA	975.	GGTTTCTTTGGGTCGCCTGGCTTTTTGGATCTCCGGGATTCTTCGGTAGTCCAGGATTTTTCGGCTCACCA
466.	GGTCACGACGGTATCCCAGGCCATGATGGAATACCCGGACATGACGGCATTCCTGGACACGATGGTATACCA	976.	GGTTTCTTTGGTACGCCCGGATTTTTCGGCACTCCTGGCTTTTTGGGACCCCTGGATTCTTCGGTACACCA
467.	GGTGGGGATCACTTACCGGGCGGAGACCATCTTCCTGGAGGAGATCATTTGCCAGGGGGTGACCACTTACCA	977.	GGTGGCTTTTTCGTGCCCGGCGGATTTTTTGTCCCTGGGGGTTTCTTTGTTCCTGGAGGATTCTTCGTACCA
468.	GGTCATGGAGACAAGCCTGGACACGGCGATAAACCCGGACATGGTGATAAGCCAGGGCATGGGGATAAACCA	978.	GGTTTTGGGTCAAGTCCGGGATTTGGATCATCGCCAGGCTTTGGTAGTTCTCCAGGATTCGGCAGTTCACCA
469.	GGTCACGGTGATAGCCCTGGGCATGGCGATAGTCCGGGTCATGGAGATTCTCCTGGACATGGGGATTCACCA	979.	GGTTCTGGCTTTACTCCAGGGAGCGGTTTTACCCCAGGCAGTGGATTCACGCCCGGATCAGGATTTACACCA
470.	GGTCACGATGGTACGCCGGGACACGACGGTACTCCAGGGCATGACGGAACCCCAGGACATGATGGCACACCA	980.	GGTGGATTTTGGAGTCCCGGAGGCTTTTGGTCTCCTGGGGGTTTCTGGAGCCCTGGCGGTTTTTGGTCACCA
471.	GGTGGGGATCATGTTCCTGGAGGAGATCACGTACCGGGCGGAGACCATGTGCCAGGGGGTGACCACGTACCA	981.	GGTTACTTCGGTTCTCCAGGCTACTTTGGCAGTCCTGGGTATTTCGGATCGCCGGGATATTTTGGTTCACCA
472.	GGTGGTGATATTATTCCTGGGGGTGACATTATACCGGGCGGAGACATAATCCCTGGAGGAGATATAATACCA	982.	GGTTTTGTTGGATCGCCAGGCTTTGTGGGGTCTCCAGGATTCGTCGGTAGTCCGGGATTTGTAGGGTCACCA
473.	GGTGGACTAGATATTCCTGGCGGTCTTGATATACCCGGAGGTTTAGACATCCCTGGGGGCTTAGATATACCA	983.	GGTACATTCGGAACTCCTGGCACATTTGGGACCCCTGGAACCTTTGGTACGCCCGGTACTTTTGGAACACCA
474.	GGTGGTGATAAAATCCCAGGGGGTGACAAAATACCGGGCGGAGATAAGATTCCTGGAGGAGACAAGATACCA	984.	GGTTGGTTCGGTACTCCAGGGTGGTTTGGCACGCCCGGCTGGTTTGGGACCCCAGGATGGTTTGGAACACCA
475.	GGTGATGGGATGATACCCGGCGATGGAATGATTCCTGGGGACGGTATGATCCCTGGAGATGGCATGATACCA	985.	GGTTACTTCGGTACGCCAGGATATTTCGGCACTCCGGGCTATTTTGGGACCCCTGGATACTTTGGTACACCA
476.	GGTGGTGACTTTATTCCTGGGGGTGATTTCATACCGGGCGGAGATTTTATCCCTGGAGGAGACTTCATACCA	986.	GGTACATTTGGAGTTCCAGGAACCTTTGGCGTACCCGGTACTTTTGGTGTGCCTGGGACATTCGGAGTACCA
477.	GGTGGAATTGATAGCCCTGGCGGTATAGACAGTCCCGGAGGCATAGATTCTCCTGGGGGTATTGACTCACCA	987.	GGTTGGTTTGGAGTGCCTGGATGGTTCGGTGTCCCTGGGTGGTTTGGCGTTCCGGGCTGGTTTGGGGTACCA
478.	GGTGGAACTGATATTCCTGGAGGTACAGACATCCCTGGGGGCACAGATATACCCGGCGGTACGGACATACCA	988.	GGTTTCTACGGCGTTCCTGGCTTTTATGGAGTCCCAGGGTTCTATGGTGTGCCCGGATTTTACGGTGTACCA
479.	GGTGGAGACTGGATACCCGGAGGAGATTGGATCCCTGGCGGTGATTGGATTCCTGGGGGCGATTGGATACCA	989.	GGTTTTGTAGGAGTGCCTGGCTTTGTGGGCGTCCCTGGATTTGTCGGTGTTCCGGGGTTCGTTGGTGTACCA
480.	GGTGGTGATTACATACCGGGCGGAGACTATATCCCTGGAGGAGATTATATTCCTGGGGGTGACTACATACCA	990.	GGTAGTGGATCAAGTCCAGGGTCAGGCAGTTCGCCAGGAAGCGGTTCATCTCCGGGCTCTGGGAGTTCACCA
481.	GGTGGAATAGATGTGCCAGGAGGAATTGACGTTCCGGGCGGTATTGATGTCCCTGGGGGTATAGACGTACCA	991.	GGTAGTACGGGAAGTCCAGGCAGCACTGGGTCTCCTGGCTCAACAGGTTCGCCCGGATCTACAGGGTCACCA
482.	GGTGATGGTCTTCTACCCGGCGATGGCTTACTTCCTGGGGACGGTTTATTGCCTGGAGATGGATTATTACCA	992.	GGTAGCGGATGGTCCCCTGGATCTGGGTGGAGTCCTGGCAGTGGTTGGTCGCCCGGTTCAGGCTGGTCACCA
483.	GGTGATGGAAAGCTCCCTGGAGACGGAAAATTACCCGGCGATGGCAAACTTCCTGGGGATGGAAATTACCA	993.	GGTGGATCATATTCTCCTGGGGGTAGTTACAGCCCTGGCGGTTCGTATAGTCCCGGAGGCAGTTATTCACCA
484.	GGTATGGATGGTTTACCCGGAATGGATGGGCTTCCTGGGATGGACGGTCTTCCAGGCATGGATGGCTTACCA	994.	GGTTCGTCAGGGGTCCCAGGAAGTAGTGGAGTTCCAGGCTCAAGCGGTGTGCCCGGATCATCTGGCGTACCA
485.	GGTTTTGGAGACTTGCCAGGGTTCGGCGATCTTCCAGGATTTGGTGATCTACCGGGCTTTGGGGACTTACCA	995.	GGTACAAGTGGTACGCCCGGTACTTCAGGGACTCCAGGAACTAGCGGAACCCCAGGCACATCTGGCACACCA
486.	GGTGATCTAGGAAGTCCGGGAGACCTCGGTAGCCCAGGGGATCTTGGCAGTCCAGGCGATTTAGGTTCACCA	996.	GGTGGAACATGGAGTCCTGGGGGTACTTGGAGCCCTGGCGGTACGTGGTCTCCCGGAGGCACTTGGTCACCA
487.	GGTGGGACAGATTTGCCAGGAGGTACTGACTTACCGGGCGGAACTGATCTTCCAGGGGGTACAGACTTACCA	997.	GGTGGATCATATACCCCTGGAGGAAGTTACACTCCCGGAGGCTCTTATACTCCTGGCGGTTCGTATACACCA
488.	GGTTGGGATGGTCTTCCAGGGTGGGATGGACTACCGGGCTGGGACGGATTGCCAGGATGGGATGGCTTACCA	998.	GGTGTCGGTACAAGTCCTGGCGTAGGCACATCTCCGGGAGTAGGTACTAGCCCTGGGGTTGGAACTTCACCA
489.	GGTGATGGATATTTGCCAGGGGACGGCTATCTACCGGGCGATGGTTATCTTCCTGGAGATGGGTATTTACCA	999.	GGTTGGGGATGGAGTCCAGGGTGGGGTTGGTCGCCAGGCTGGGGGTGGTCTCCCGGATGGGGCTGGTCACCA
490.	GGTTTGGGCGATGTACCCGGTCTAGGTGATGTTCCAGGACTCGGAGATGTGCCTGGCTTAGGGGATGTACCA	1000.	GGTTGGTATGGATCTCCCGGCTGGTATGGTTCGCCAGGATGGTACGGAAGTCCAGGGTGGTATGGGTCACCA
491.	GGTAAAGGTGATAAGCCAGGAAAAGGAGACAAGCCTGGGAAAGGCGATAAACCGGGCAAAGGGGATAAACCA	1001.	GGTTGGGTCGGTTCTCCAGGGTGGGTTGGAAGTCCGGGATGGGTAGGATCGCCAGGCTGGGTAGGTTCACCA
492.	GGTGGTATGGATAAACCCGGCGGAATGGACAAGCCAGGAGGGATGGATAAGCCTGGGGGCATGGATAAACCA	1002.	GGTGGATATTATAGCCCTGGGGGTTATTACAGTCCTGGCGGTTACTATTCTCCGGGAGGATACTACTCACCA
493.	GGTTTTGACGGCAAGCCTGGGTTCGATGGTAAACCGGGATTTGATGGAAAGCCAGGATTCGACGGAAAACCA	1003.	GGTTATTCTGGAGTTCCTGGGTATTCAGGCGTGCCCGGATATAGTGGTGTCCCTGGCTACAGCGGTGTACCA
494.	GGTGGCAAGGATTCTCCAGGGGGTAAGGACAGCCCAGGCGGTAAAGATAGTCCGGGAGGAAAAGACTCACCA	1004.	GGTGGAGTTGTTAGCCCAGGCGGTGTCGTAAGTCCAGGAGGGGTAGTATCTCCCGGAGGAGTAGTGTCACCA
495.	GGTAAGGACGGAACCCCAGGAAAAGATGGCACTCCAGGGAAGGATGGTACGCCCGGTAAAGACGGCACACCA	1005.	GGTACTACTGGTACGCCTGGGACGACCGGAACTCCTGGCACTACAGGAACCCCTGGAACAACGGGCACACCA
496.	GGTGGATGGGATAAACCGGGCGGCTGGGATAAGCCTGGAGGTTGGGACAAGCCAGGAGGGTGGGATAAACCA	1006.	GGTACGTGGGGTACTCCCGGTACATGGGGAACGCCAGGAACTTGGGGCACTCCAGGGACCTGGGGGACACCA
497.	GGTGGAGATTACAAACCGGGCGGAGACTACAAGCCAGGAGGGGACTATAAGCCTGGGGGTGATTATAAACCA	1007.	GGTTATACGGGAACCCCTGGATACACAGGCACGCCCGGATATACAGGTACTCCTGGCTATACTGGGACACCA
498.	GGTGATGGTAAAGTCCCTGGGGATGGAAAAGTACCGGGCGATGGCAAAGTTCCTGGAGACGGAAAGGTACCA	1008.	GGTGTAACTGGCACTCCCGGAGTCACAGGAACTCCTGGGGTTACTGGTACTCCAGGCGTAACAGGGACACCA
499.	GGTGATATGGGAAGTCCAGGGGATATGGGCAGTCCCGGAGACATGGGTAGTCCTGGCGATATGGGGTCACCA	1009.	GGTGGGTGGTGGACCCCTGGAGGCTGGTGGACTCCTGGCGGATGGTGGACACCGGGAGGTTGGTGGACACCA
500.	GGTGACATGGGCACTCCAGGGGATATGGGTACTCCTGGCGATATGGGAACTCCGGGAGATATGGGGACACCA	1010.	GGTTGGTATGGCACGCCGGGCTGGTATGGACCCCTGGGTGGTATGGAACTCCTGGATGGTACGGTACACCA
501.	GGTGGGATGGATGTTCCTGGCGGTATGGATGTACCCGGAGGAATGGACGTTCCAGGGGGCATGGATGTACCA	1011.	GGTTGGGTAGGGACGCCTGGATGGGTAGGAACTCCGGGGTGGGTCGGCACTCCTGGCTGGGTTGGTACACCA
502.	GGTGGATTTGATAGTCCGGGAGGATTCGACAGCCCTGGGGGTTTTGACTCTCCTGGCGGTTTCGATTCACCA	1012.	GGTTATTATGGAACGCCAGGATATTACGGCACTCCGGGCTACTATGGGACCCCTGGATACTACGGTACACCA
503.	GGTTTTGATGGGACCCCTGGATTTGACGGCACTCCTGGCTTCGATGGTACGCCCGGATTCGACGGTACACCA	1013.	GGTTATGTGGGCACGCCTGGCTATGTCGGTACTCCCGGATATGTAGGAACCCCTGGATACGTTGGTACACCA
504.	GGTGGATTTGACGTTCCGGGCGGTTTTGATGTGCCAGGAGGCTTCGATGTCCCTGGGGGTTTCGACGTACCA	1014.	GGTGGAACAGTCGTTCCTGGAGGTACGGTTGTCCCAGGGGGCACTGTAGTACCCGGCGGAACTGTGGTACCA

505. GGTGATTCAGGCAGTCCGGGAGACAGCGGTAGCCCAGGGGATTCTGGAAGTCCTGGCGATAGTGGTTCACCA	1015. GGTTGGTGGGGTGTGCCAGGGTGGTGGGGCGTACCCGGATGGTGGGGAGTTCCAGGCTGGTGGGGGGTACCA
506. GGTAGTGATGGGACCCCTGGCTCTGACGGCACGCCCGGTTCGGATGGAACTCCAGGAAGCGATGGTACACCA	1016. GGTGGGTGGTACGTTCCGGGAGGATGGTATGTTCCAGGCGGTTGGTATGTGCCAGGGGGCTGGTATGTACCA
507. GGTGATGGGTGGTCTCCGGGAGATGGTTGGAGCCCTGGGGACGGCTGGAGTCCTGGCGATGGATGGTCACCA	1017. GGTGTAGGGTGGGTCCCCGGAGTCGGCTGGGTTCCTGGAGTAGGATGGGTGCCTGGCGTAGGTTGGGTACCA
508. GGTGGATATGACAGTCCGGGAGGATACGATTCTCCTGGCGGTTATGATAGCCCTGGGGGTTACGACTCACCA	1018. GGTGGATACTACGTTCCTGGCGGTTACTATGTCCCTGGGGGTTATTATGTGCCCGGAGGCTATTACGTACCA
509. GGTGATGTTGGCAGCCCTGGAGATGTAGGAAGTCCGGGTGACGTAGGTTCGCCTGGGGATGTCGGTTCACCA	1019. GGTGGATATGTTGTGCCAGGAGGCTATGTCGTCCCAGGGGGTTATGTAGTTCCGGGCGGTTACGTTGTACCA
510. GGTACGGGAGATACTCCTGGGACTGGCGATACACCCGGAACAGGTGACACTCCAGGCACAGGGGATACACCA	1020. GGTGTAGTAGGAGTTCCAGGCGTAGTTGGAGTACCGGGAGTCGTCGGAGTGCCAGGGGTTGTGGGCGTACCA

Table 12: Summary of alignments of next generation sequencing (NGS) data grouped by oligonucleotide reference ID and concise idiosyncratic gapped alignment report (CIGAR)

Oligo ID	Fwd/ Rev	CIGAR string alignment	Occurrences in NGS
7	Fwd	1X	2
7	Fwd	=	50
9	Fwd	1X	355
9	Fwd	1X3=1X	1
9	Fwd	=	507
10	Fwd	1D28=1I24=1X	248
10	Fwd	1X	1
10	Fwd	=	160
13	Fwd	1I5=1D	44
13	Fwd	1X	2
13	Fwd	3D	55
13	Fwd	=	118
14	Fwd	1I25=1X3=1D	6
14	Fwd	=	4
15	Fwd	1I13=1D	118
15	Fwd	1I6=1X6=1D	1
15	Fwd	1X	13
15	Fwd	=	155
22	Fwd	1X	1
22	Fwd	=	131
25	Fwd	1D6=1I	138
27	Fwd	1D12=2D32=1X	20
27	Fwd	1X	58
27	Fwd	1X11=1X	7
27	Fwd	=	239
28	Fwd	1X	1
28	Fwd	=	281
33	Fwd	1X	1
33	Fwd	=	62
34	Fwd	1D11=1I	42
34	Fwd	1I12=1D	43
36	Fwd	=	139
41	Fwd	1D6=1I	186
41	Fwd	1X	12
41	Fwd	3D	24
41	Fwd	=	218
42	Fwd	1X	137
46	Fwd	=	27
48	Fwd	1X	3
48	Fwd	1X3=2X	41
50	Fwd	1X14=1X	66
51	Fwd	2I11=2D	93
51	Fwd	3X9=6D	1
51	Fwd	=	224
55	Fwd	1X	117
59	Fwd	1X	1
60	Fwd	=	14
61	Fwd	=	29
65	Fwd	1D12=1D1X17=1D	3
65	Fwd	1D14=2D	42
65	Fwd	1D17=1I	5
65	Fwd	1D23=1I	109
65	Fwd	1D37=1I	45
65	Fwd	1D38=1I	41
65	Fwd	1D4=1I	93
65	Fwd	1D7=1I	133
65	Fwd	1D8=1I	119
65	Fwd	1I16=1I4=1I	1
65	Fwd	1I2=1I9=1I	1
65	Fwd	1I30=2D8=1I	71
65	Fwd	1I35=1D	73
65	Fwd	1I39=1D	45
65	Fwd	1I44=1D	82
65	Fwd	1I5=1D	96
65	Fwd	1I6=1D	17
65	Fwd	1X	7225
65	Fwd	1X16=1X	1
65	Fwd	1X17=1X	1
65	Fwd	1X21=1X	20
65	Fwd	1X29=1X	87
65	Fwd	1X2=1X	610
65	Fwd	1X35=1X	862
65	Fwd	2X	44
65	Fwd	3D	229
65	Fwd	6D	562
65	Fwd	=	70360
67	Fwd	=	91
70	Fwd	1D10=1I9=1X	43
70	Fwd	1D5=1I	168
70	Fwd	1X	51
70	Fwd	3D	2
70	Fwd	=	345
71	Fwd	1X	272
71	Fwd	1X1=1X5=3D9=1X	103
71	Fwd	=	188
73	Fwd	1X2=1D14=1I	3
73	Fwd	=	166
74	Fwd	1D13=1I	2
79	Fwd	1D37=1I	78
83	Fwd	=	7
85	Fwd	1X	1
85	Fwd	=	69
93	Fwd	=	40
95	Fwd	1X	1
95	Fwd	=	52
96	Fwd	=	4
97	Fwd	1I23=1D	2
97	Fwd	2I12=2D	10
97	Fwd	=	1
100	Fwd	3D	485
100	Fwd	=	37
103	Fwd	1X42=1X	7
104	Fwd	1I33=1I4=1I	5
104	Fwd	1X	144
106	Fwd	1X	270
106	Fwd	=	225
108	Fwd	=	40
109	Fwd	1I8=1I2=1X30=1I	143
109	Fwd	1X	1981
109	Fwd	1X2=1X	8671
109	Fwd	1X37=1X2=1X	5
109	Fwd	2X	1
109	Fwd	3D	38
109	Fwd	=	10522
111	Fwd	1D11=1I	159
111	Fwd	=	49
112	Fwd	1D8=1I	32
115	Fwd	1X	2682
115	Fwd	1X9=1X	1
115	Fwd	3D	1085
115	Fwd	=	8453
119	Fwd	=	32
120	Fwd	1X	2
120	Fwd	=	79
121	Fwd	=	98
123	Fwd	1X	30
123	Fwd	2X	43
123	Fwd	=	190
127	Fwd	=	42
129	Fwd	1I13=1I5=1I	5
129	Fwd	1I6=1D	119
129	Fwd	1X	288
129	Fwd	1X24=1X	67
129	Fwd	=	4414
131	Fwd	1I7=1D	49
131	Fwd	1X	128
131	Fwd	=	80
133	Fwd	=	89
134	Fwd	1X	21
134	Fwd	=	299
137	Fwd	=	13
138	Fwd	1X	6
140	Fwd	1I7=1D	16
140	Fwd	1X	150
140	Fwd	=	303
141	Fwd	=	1
145	Fwd	2D23=1I2=1I	17
145	Fwd	=	62
147	Fwd	=	69
148	Fwd	=	157
149	Fwd	1I7=2X3=1D	104
150	Fwd	1X	20
150	Fwd	=	65
154	Fwd	1X	168
154	Fwd	1X2=1I36=1D	1
154	Fwd	=	87
155	Fwd	1X	2
155	Fwd	1X16=1X	14
155	Fwd	=	249
156	Fwd	=	8
162	Fwd	=	94
163	Fwd	1X	2
163	Fwd	=	70
164	Fwd	=	206
165	Fwd	1X	6
165	Fwd	=	218
167	Fwd	1X	1
167	Fwd	=	32
168	Fwd	1D16=1I	11
168	Fwd	1D20=1I	19
168	Fwd	1D33=1I11=1X	26
168	Fwd	1X	5
168	Fwd	=	238
169	Fwd	1I26=1D	2
169	Fwd	1I32=1D	22
169	Fwd	1I34=1D	12
169	Fwd	1I6=4D	10
169	Fwd	1X	385
169	Fwd	1X6=1D34=1I	121
169	Fwd	2D1=1X45=1D	71
169	Fwd	3D	505
169	Fwd	=	1287
170	Fwd	1D16=1I	58
170	Fwd	1D5=1I	39
170	Fwd	1I10=1D	50
170	Fwd	1I27=1I21=1I	29
170	Fwd	1I31=1D	39
170	Fwd	1I43=1D	6
170	Fwd	1I6=1D	36
170	Fwd	1I6=1D1=1X	175
170	Fwd	1I9=1D	187
170	Fwd	1X	639
170	Fwd	1X20=1X	144
170	Fwd	1X38=1X	105
170	Fwd	2X3=3I	27
170	Fwd	3D	84
170	Fwd	3X5=1D21=1I14=1X	82
170	Fwd	=	4508
182	Fwd	1X	13
184	Fwd	1D27=1I	11
184	Fwd	1I11=4D	68
184	Fwd	=	175
185	Fwd	=	95
188	Fwd	1X	1
188	Fwd	=	35
190	Fwd	1D43=1I	17
191	Fwd	=	29
192	Fwd	1D4=1X15=2D2=1X	2
194	Fwd	1D18=1I	18
194	Fwd	1I11=1I10=1I	15
194	Fwd	1I4=1D	34
194	Fwd	1X	147
194	Fwd	=	3402
195	Fwd	1X	1
196	Fwd	1X	6
196	Fwd	1X11=1X	1
196	Fwd	=	66
197	Fwd	3D1X	41
201	Fwd	1I4=3D10=1D	1
201	Fwd	1X	109
201	Fwd	=	232
205	Fwd	1X	342
214	Fwd	=	66
225	Fwd	=	2
229	Fwd	1D21=1I	46
230	Fwd	1I39=1D	38
230	Fwd	=	99
239	Fwd	=	112
249	Fwd	=	5

259	Fwd	1I9=4D	44
259	Fwd	1X	188
259	Fwd	=	111
260	Fwd	1X	71
261	Fwd	1I29=1D	4
261	Fwd	1X20=1X	2
261	Fwd	=	32
262	Fwd	1X1=1X	79
266	Fwd	1X	42
267	Fwd	1D25=1I	5
272	Fwd	=	47
276	Fwd	1I10=1D15=1I7=1D	1
282	Fwd	=	166
283	Fwd	=	53
287	Fwd	1X	2
287	Fwd	=	31
293	Fwd	1X	5
293	Fwd	=	31
296	Fwd	=	75
300	Fwd	=	4
303	Fwd	=	28
304	Fwd	=	7
305	Fwd	1X15=1X	8
308	Fwd	1X	43
310	Fwd	1X	2
310	Fwd	2X2=1X	1
310	Fwd	=	21
313	Fwd	1D18=1I	11
313	Fwd	2D4=1X6=1D	29
314	Fwd	1I15=1D	3
314	Fwd	1X	53
314	Fwd	3D1X	57
314	Fwd	=	227
315	Fwd	1X	99
315	Fwd	=	49
316	Fwd	1X	28
317	Fwd	3D	40
317	Fwd	6D	2
317	Fwd	=	249
320	Fwd	1X	55
320	Fwd	=	37
322	Fwd	1I1X27=1D	42
322	Fwd	1I8=1X9=1I9=1I	8
322	Fwd	1X	74
322	Fwd	1X9=1X	18
325	Fwd	=	6
327	Fwd	1X	361
327	Fwd	1X25=1X	5
327	Fwd	6D	33
327	Fwd	=	829
328	Fwd	1X	123
328	Fwd	=	233
331	Fwd	1I17=1D	152
331	Fwd	1X	112
331	Fwd	=	539
332	Fwd	1X	251
332	Fwd	=	26
338	Fwd	1X	232
338	Fwd	=	388
341	Fwd	1X	94
346	Fwd	=	168
350	Fwd	1X	9
350	Fwd	2I32=1I	10
350	Fwd	3D	43
358	Fwd	1I6=1I19=2D	9
358	Fwd	1X	223
358	Fwd	1X1=2X	27
358	Fwd	=	1309

368	Fwd	=	1
374	Fwd	1D12=1I	24
374	Fwd	1I33=1D	18
374	Fwd	1X	748
374	Fwd	1X14=1X	3
374	Fwd	1X19=1X	56
374	Fwd	3D	257
374	Fwd	=	3274
375	Fwd	1I30=7D	12
375	Fwd	1X	98
376	Fwd	=	74
377	Fwd	1X	1
381	Fwd	1X	78
382	Fwd	=	13
384	Fwd	=	3
386	Fwd	1X	67
386	Fwd	1X18=1X	1
386	Fwd	=	64
387	Fwd	1I4=1D	21
395	Fwd	1X	50
397	Fwd	=	25
398	Fwd	1X	1
398	Fwd	=	96
402	Fwd	1X	4
402	Fwd	=	128
407	Fwd	1X	46
407	Fwd	=	20
413	Fwd	1X1=1X	16
414	Fwd	=	21
416	Fwd	1X	337
416	Fwd	=	161
417	Fwd	1X	82
417	Fwd	1X21=1X	88
417	Fwd	2D10=1D	62
420	Fwd	1X	236
421	Fwd	=	13
429	Fwd	=	15
433	Fwd	1I8=1I26=1I	5
435	Fwd	1X	17
435	Fwd	1X24=1X	22
443	Fwd	1X	4
443	Fwd	=	49
446	Fwd	1X	2
446	Fwd	=	87
450	Fwd	1D9=1I	26
450	Fwd	1I20=1I7=1I	53
450	Fwd	1X	108
452	Fwd	=	64
454	Fwd	1X	121
455	Fwd	1D23=1I	33
467	Fwd	1D47=1X5=1I	7
467	Fwd	=	127
468	Fwd	1D9=1I	47
468	Fwd	=	1
469	Fwd	=	211
470	Fwd	=	117
471	Fwd	1X	6
471	Fwd	=	151
472	Fwd	=	75
473	Fwd	1X	130
473	Fwd	=	59
474	Fwd	1X	10
474	Fwd	=	241
477	Fwd	=	8
478	Fwd	1D10=1I	2
478	Fwd	1D21=1I	108
478	Fwd	1X	537
478	Fwd	1X19=1X	127
478	Fwd	1X24=1X	3

478	Fwd	2I19=1I	1
478	Fwd	4D9=1I	2
478	Fwd	=	161
479	Fwd	1X34=1X7=1X	38
480	Fwd	1I15=1D	4
480	Fwd	1I25=1D	42
480	Fwd	1I38=1D	5
480	Fwd	2X	1
482	Fwd	=	56
484	Fwd	=	4
489	Fwd	=	117
490	Fwd	=	6
491	Fwd	3X	56
495	Fwd	=	49
498	Fwd	=	53
499	Fwd	1X	10
499	Fwd	=	98
500	Fwd	1X	450
500	Fwd	=	121
502	Fwd	=	40
503	Fwd	1X	45
507	Fwd	=	57
510	Fwd	1D32=1X8=1I	217
510	Fwd	6D	300
511	Fwd	=	100
512	Fwd	1X	68
517	Fwd	1I20=1D17=1X	70
517	Fwd	=	173
520	Fwd	1X	134
520	Fwd	=	542
522	Fwd	1X	81
522	Fwd	=	355
525	Fwd	=	50
527	Fwd	=	54
528	Fwd	1X11=1X	3
528	Fwd	=	24
546	Fwd	=	3
547	Fwd	1I10=1D	20
548	Fwd	=	24
557	Fwd	1I14=1D	5
557	Fwd	1I18=1D24=1X	59
557	Fwd	1I32=2I	3
557	Fwd	1I9=1D	3
557	Fwd	1X10=1X	1
557	Fwd	1X18=1X	25
557	Fwd	1X1=1X	82
557	Fwd	3D	36
557	Fwd	3I	7
557	Fwd	3X	1
557	Fwd	6D	1
557	Fwd	=	1717
558	Fwd	1X	17
558	Fwd	=	183
562	Fwd	=	1
563	Fwd	1D11=2D	230
563	Fwd	1X	2
563	Fwd	3D4=3D	4
563	Fwd	=	181
564	Fwd	1X	1
565	Fwd	1I13=1D	19
565	Fwd	1I8=1D	72
565	Fwd	1X	778
565	Fwd	3D	542
565	Fwd	4D1=1X26=2D	9
565	Fwd	=	975
566	Fwd	1X	76
566	Fwd	=	22

570	Fwd	1X19=1X5=1X	1
570	Fwd	=	38
572	Fwd	1I21=1I9=1I	3
572	Fwd	1X	23
572	Fwd	=	148
575	Fwd	1X	4
575	Fwd	=	43
577	Fwd	1D16=1I	12
577	Fwd	1I33=1D	48
577	Fwd	1X	131
577	Fwd	=	803
581	Fwd	1X	51
581	Fwd	=	448
582	Fwd	=	38
583	Fwd	1X	5
583	Fwd	=	130
584	Fwd	1X	114
584	Fwd	1X18=1X	29
588	Fwd	=	37
592	Fwd	1D6=2D	24
592	Fwd	=	3
595	Fwd	1D31=1I	12
595	Fwd	1D6=1I19=1X5=1X	26
595	Fwd	1I18=1X23=1D	1
595	Fwd	1I24=4D	21
595	Fwd	1I26=1D	388
595	Fwd	1I31=1D10=1X	4
595	Fwd	=	2580
597	Fwd	=	9
599	Fwd	1X	2
599	Fwd	=	82
600	Fwd	1I3=1X5=1I4=1I	1
600	Fwd	1X	136
600	Fwd	=	58
603	Fwd	1X	1
603	Fwd	=	61
606	Fwd	1X	57
608	Fwd	1X	8
612	Fwd	1D36=1I	21
612	Fwd	1X	4
612	Fwd	=	90
619	Fwd	1X	8
619	Fwd	=	16
620	Fwd	1X	2
621	Fwd	=	34
623	Fwd	=	18
631	Fwd	1X	2
631	Fwd	=	224
643	Fwd	2D3=1D	53
647	Fwd	1X	61
648	Fwd	1X	5
648	Fwd	3D	18
648	Fwd	=	355
649	Fwd	=	83
651	Fwd	1X	78
651	Fwd	=	163
653	Fwd	1X	26
653	Fwd	=	377
658	Fwd	1X	3
658	Fwd	=	156
659	Fwd	1X	28
660	Fwd	1D8=1I	95
660	Fwd	1X20=1X	3
660	Fwd	=	1
666	Fwd	6D	21
666	Fwd	=	35
667	Fwd	=	48

668	Fwd	1D39=2D	4
668	Fwd	1I18=1D	86
668	Fwd	1X	644
668	Fwd	=	3253
675	Fwd	1X	105
676	Fwd	1X	85
684	Fwd	1X	2
684	Fwd	=	170
690	Fwd	=	6
693	Fwd	=	75
697	Fwd	1X	1
700	Fwd	1X	79
700	Fwd	=	72
701	Fwd	1D33=2D	25
701	Fwd	=	60
702	Fwd	1X	87
705	Fwd	1D10=2D	10
710	Fwd	1X	3
711	Fwd	1X13=1D6=1I	2
711	Fwd	=	131
719	Fwd	1X	9
719	Fwd	=	1
720	Fwd	1X	2
720	Fwd	=	124
725	Fwd	1X28=1X	2
728	Fwd	1D51=1I	43
728	Fwd	1I24=1D	146
728	Fwd	=	3
729	Fwd	=	13
731	Fwd	=	43
733	Fwd	=	1
739	Fwd	1X	94
740	Fwd	1X	1
742	Fwd	=	18
752	Fwd	1D36=1I	5
752	Fwd	1I17=1I9=1I	1
752	Fwd	1X	225
752	Fwd	1X23=1X	11
752	Fwd	2D2=2I1=1X	1
752	Fwd	=	738
753	Fwd	1D42=1I	3
753	Fwd	1X	8
753	Fwd	6D	5
753	Fwd	=	247
754	Fwd	2X	2
754	Fwd	=	2
756	Fwd	1D25=1I	2
756	Fwd	=	2
764	Fwd	1X	4
765	Fwd	1X	126
765	Fwd	1X6=1X	1
765	Fwd	3D	1
765	Fwd	=	3873
771	Fwd	1I36=1D	82
771	Fwd	1X	203
771	Fwd	1X17=1X	3
771	Fwd	=	352
772	Fwd	1X	401
772	Fwd	3X	3
772	Fwd	6D	61
772	Fwd	=	788
773	Fwd	1D34=1I	199
773	Fwd	1X	56
773	Fwd	=	50
774	Fwd	1X	1
774	Fwd	=	87
777	Fwd	1X	1
777	Fwd	3D	2
777	Fwd	=	18

778	Fwd	3D	2
778	Fwd	6D	70
778	Fwd	=	174
779	Fwd	=	106
780	Fwd	1X	17
780	Fwd	=	209
781	Fwd	1X3=1X	1
781	Fwd	=	250
782	Fwd	1X	18
782	Fwd	=	136
783	Fwd	=	142
784	Fwd	1D16=7D25=1D	167
784	Fwd	1I30=1D	62
784	Fwd	1X	119
784	Fwd	1X24=1D7=1I	29
785	Fwd	1X	127
785	Fwd	1X2=1D17=1I	1
785	Fwd	=	301
786	Fwd	1D14=1I	2
786	Fwd	1D37=1I	233
786	Fwd	1I46=1D	4
786	Fwd	1X	130
786	Fwd	1X28=3D	1
786	Fwd	1X2=1X	4
786	Fwd	3D	3
786	Fwd	3D25=1X2=2X	72
786	Fwd	=	806
788	Fwd	1X	1
788	Fwd	=	4
789	Fwd	=	25
790	Fwd	1X	3
790	Fwd	=	89
791	Fwd	1D22=1I8=1X	100
791	Fwd	1D8=1I	71
791	Fwd	1I7=1I45=1I	31
791	Fwd	1X	321
791	Fwd	1X48=1X	33
791	Fwd	2I48=1I	25
791	Fwd	3D	80
791	Fwd	=	3214
792	Fwd	1X	8
792	Fwd	1X23=1X	1
792	Fwd	=	160
793	Fwd	1X	1
793	Fwd	=	33
794	Fwd	1X	8
794	Fwd	=	9
797	Fwd	1D21=1I	46
798	Fwd	1X	71
798	Fwd	1X15=1X	6
798	Fwd	=	182
799	Fwd	1I2=1I27=1I	39
799	Fwd	=	33
801	Fwd	1D10=1I	7
803	Fwd	1X	1
803	Fwd	=	55
804	Fwd	1X	67
804	Fwd	=	32
805	Fwd	1X	21
806	Fwd	1I13=1D	6
806	Fwd	1X	177
806	Fwd	=	307
808	Fwd	1X	141
808	Fwd	=	93
809	Fwd	1X12=3D	114

809	Fwd	1X1=1X	62
809	Fwd	=	48
810	Fwd	=	102
811	Fwd	1I38=1D	7
811	Fwd	1X	55
811	Fwd	1X19=1X	41
811	Fwd	1X3=1X	169
811	Fwd	3D	8
811	Fwd	=	667
813	Fwd	1D21=1I	16
813	Fwd	1X	169
813	Fwd	1X6=1X	1
813	Fwd	=	855
814	Fwd	1I26=2D4=1I	65
814	Fwd	=	145
815	Fwd	1D20=1I	5
816	Fwd	1X	28
816	Fwd	1X41=1X	1
817	Fwd	=	13
818	Fwd	1I6=1I34=1I	12
820	Fwd	1I25=1D	12
820	Fwd	1X	104
820	Fwd	2I1X24=1I	26
820	Fwd	=	584
821	Fwd	1I28=1D	57
821	Fwd	1X	1
821	Fwd	1X30=1D8=1I	142
821	Fwd	3D	144
821	Fwd	=	190
823	Fwd	=	24
828	Fwd	1D4=1I	30
828	Fwd	1X	76
828	Fwd	2X	3
828	Fwd	=	72
829	Fwd	1I16=1I9=1I	115
829	Fwd	1X	2
829	Fwd	=	323
830	Fwd	1I12=1I7=1I	42
830	Fwd	1X	1
830	Fwd	=	177
834	Fwd	=	74
835	Fwd	1X	2
835	Fwd	3D2=1X	10
835	Fwd	=	59
837	Fwd	=	2
838	Fwd	2X14=9D	1
838	Fwd	2X33=1X	2
838	Fwd	3D	21
839	Fwd	1X	1
839	Fwd	=	41
840	Fwd	=	61
843	Fwd	1X	2
846	Fwd	=	31
848	Fwd	1X	7
849	Fwd	1X	80
851	Fwd	=	50
853	Fwd	1D7=1I	29
853	Fwd	1I38=1D	76
853	Fwd	1X	122
853	Fwd	=	157
854	Fwd	1X	77
856	Fwd	1X	1
856	Fwd	=	87
861	Fwd	1X	56
862	Fwd	1X	344
863	Fwd	1X	8
863	Fwd	=	1

865	Fwd	=	15
868	Fwd	3D	69
868	Fwd	=	555
869	Fwd	1I39=1D	47
871	Fwd	1D7=1I	9
871	Fwd	1I1=1X39=1I5=1I	18
871	Fwd	1X	19
871	Fwd	1X38=1X	52
871	Fwd	=	52
872	Fwd	=	16
873	Fwd	2D8=1X7=1D	2
873	Fwd	=	65
874	Fwd	1D13=2D	5
874	Fwd	1D20=1I24=1X	3
874	Fwd	1X	1
874	Fwd	=	120
875	Fwd	1X	15
875	Fwd	=	196
878	Fwd	1X	2
878	Fwd	=	122
882	Fwd	1X	32
883	Fwd	1D20=1I	1
883	Fwd	1X	70
883	Fwd	=	32
884	Fwd	1X	1
884	Fwd	=	8
894	Fwd	=	16
895	Fwd	1X	1
895	Fwd	=	110
898	Fwd	1X	99
898	Fwd	2X	52
898	Fwd	=	402
899	Fwd	1X	81
899	Fwd	1X7=1X	68
899	Fwd	=	67
900	Fwd	=	5
902	Fwd	=	1
906	Fwd	1X	6
906	Fwd	=	456
907	Fwd	=	210
911	Fwd	1I11=1D36=1X	5
911	Fwd	1X	11
911	Fwd	=	93
914	Fwd	1D19=1I	79
914	Fwd	1X13=1X	7
914	Fwd	=	144
916	Fwd	1X	25
918	Fwd	1D6=2D22=1X	5
918	Fwd	=	55
919	Fwd	1X2=1X	26
919	Fwd	=	55
920	Fwd	1X	48
920	Fwd	=	6
922	Fwd	1X	1
922	Fwd	=	18
932	Fwd	1X	27
932	Fwd	=	479
934	Fwd	1D23=1I	24
934	Fwd	1X	110
934	Fwd	=	1169
937	Fwd	2X1=1X29=1X	27
938	Fwd	=	114
945	Fwd	1X	2
945	Fwd	=	57
946	Fwd	=	4
949	Fwd	1X	39

949	Fwd	1X1=1X	107
949	Fwd	=	207
956	Fwd	1X	1
956	Fwd	=	31
958	Fwd	=	377
960	Fwd	=	35
961	Fwd	1D19=1I7=1X	41
961	Fwd	1I24=1D	12
961	Fwd	1X	68
961	Fwd	=	256
963	Fwd	1X	159
963	Fwd	1X5=1I8=1D	29
963	Fwd	3D	46
963	Fwd	=	550
964	Fwd	6D1X	3
966	Fwd	1X	85
966	Fwd	1X29=1X	10
966	Fwd	=	35
967	Fwd	=	56
968	Fwd	=	24
969	Fwd	1X	1
970	Fwd	1D22=1I	21
970	Fwd	1I31=1D	72
970	Fwd	1I4=1I11=1I19=1X	11
970	Fwd	1X	317
970	Fwd	3D	40
970	Fwd	=	910
977	Fwd	=	2
980	Fwd	=	159
981	Fwd	1D22=2D	38
981	Fwd	1X	138
981	Fwd	=	184
983	Fwd	1X	5
983	Fwd	=	219
986	Fwd	1D29=1I	61
986	Fwd	1D36=1I	2
986	Fwd	1I21=1I9=1I	1
986	Fwd	1X	128
986	Fwd	2I22=1I17=3D	11
986	Fwd	2X	71
986	Fwd	=	44
988	Fwd	=	15
989	Fwd	1X	1
989	Fwd	=	36
990	Fwd	1X9=1X	79
990	Fwd	=	52
991	Fwd	1D22=1I	77
991	Fwd	1I16=1I20=1I	84
991	Fwd	1X	168
991	Fwd	=	338
993	Fwd	1D15=1I	8
993	Fwd	1X	35
993	Fwd	4D11=1I	1
993	Fwd	=	16
995	Fwd	1I9=1D	47
995	Fwd	=	192
996	Fwd	1X	69
997	Fwd	1I20=1X4=1D12=1X	52
997	Fwd	1I36=1I2=1I	23
997	Fwd	1X	333
997	Fwd	1X5=1I38=1D	37
997	Fwd	=	1553
998	Fwd	=	64
1002	Fwd	1I20=1D	57

1002	Fwd	1X	1
1002	Fwd	=	28
1003	Fwd	1X	134
1003	Fwd	1X14=1X	21
1003	Fwd	=	636
1005	Fwd	=	33
1007	Fwd	1I38=1D	9
1007	Fwd	3D	40
1007	Fwd	3D1X	1730
1009	Fwd	=	6
1012	Fwd	1I7=1D	27
1012	Fwd	=	44
1013	Fwd	1D5=1I33=1X	124
1013	Fwd	1I12=1D	139
1013	Fwd	1I17=1D	7
1013	Fwd	1I26=2I	69
1013	Fwd	1I5=1D	11
1013	Fwd	1I8=1D	22
1013	Fwd	1X	5221
1013	Fwd	1X11=1X	6
1013	Fwd	1X52=1X	254
1013	Fwd	1X9=1X	10
1013	Fwd	2X	123
1013	Fwd	4X	21
1013	Fwd	=	25973
1018	Fwd	1X	21
1018	Fwd	1X18=1X	1
1018	Fwd	2I3=1I	8
1018	Fwd	2X	8
1018	Fwd	=	38
1019	Fwd	=	30
1020	Fwd	1X	5
1020	Fwd	1X8=1X	3
1020	Fwd	=	16
9	Rev	1X	11
9	Rev	=	147
10	Rev	1D28=1I24=1X	81
12	Rev	=	1
27	Rev	1X11=1X	11
28	Rev	1X	48
28	Rev	=	44
29	Rev	1X	2
32	Rev	1X	3
32	Rev	=	158
36	Rev	1D3=1I	14
37	Rev	=	39
48	Rev	1X	1
48	Rev	=	66
51	Rev	1X	1
51	Rev	=	2
52	Rev	1X	1
52	Rev	=	194
100	Rev	3D	4
109	Rev	1X2=1X	761
109	Rev	1X3=3I	77
109	Rev	6D	54
109	Rev	=	1934
120	Rev	=	112
129	Rev	1X	195
129	Rev	=	662
140	Rev	1X	150
140	Rev	=	86
144	Rev	1X	1
144	Rev	=	29
150	Rev	=	35
165	Rev	1D12=1I	3
165	Rev	1X	1
168	Rev	1D20=1I	29
168	Rev	1X	43
168	Rev	=	1

171	Rev	1X	118
182	Rev	1D19=1I	23
184	Rev	1I11=4D	11
184	Rev	=	108
201	Rev	=	2
207	Rev	1X	1
207	Rev	=	9
215	Rev	=	97
216	Rev	=	69
276	Rev	1X	104
276	Rev	=	59
279	Rev	=	2
292	Rev	1X	1
293	Rev	=	41
314	Rev	1X	30
314	Rev	=	1
320	Rev	=	61
327	Rev	1I10=1I16=1I	9
327	Rev	1X	51
328	Rev	=	23
331	Rev	1X	4
331	Rev	=	45
332	Rev	1X	1
335	Rev	=	109
359	Rev	=	65
368	Rev	=	2
372	Rev	=	1
386	Rev	1X	6
386	Rev	=	1
395	Rev	=	1
407	Rev	1I3=1D	7
412	Rev	=	10
417	Rev	1X	4
417	Rev	=	53
419	Rev	1X	1
419	Rev	=	77
431	Rev	=	1
450	Rev	1X	45
450	Rev	2I3=1X42=1I	72
476	Rev	1X	74
478	Rev	1D10=1I	1
478	Rev	1X	114
478	Rev	=	4
482	Rev	1X	31
486	Rev	1I35=1D	7
486	Rev	1X	5
486	Rev	=	102
494	Rev	1X	2
494	Rev	=	2
499	Rev	1X	4
500	Rev	1D35=1I17=1X	8
500	Rev	1X	63
500	Rev	3D28=1X	5
500	Rev	=	66
510	Rev	6D	3
557	Rev	1I30=1D1X	5
557	Rev	1X	155
565	Rev	1I44=4D	2
595	Rev	1X	65
595	Rev	=	391
631	Rev	=	9
643	Rev	=	10
648	Rev	1X	38
655	Rev	=	1
657	Rev	1X	70
657	Rev	=	4
660	Rev	1I2=1I23=1I	1
660	Rev	1X	2
660	Rev	=	74
666	Rev	1I14=1D	1
666	Rev	1X	1
666	Rev	=	102
667	Rev	=	1

668	Rev	1X	28
668	Rev	=	773
672	Rev	=	27
673	Rev	=	24
680	Rev	1X	18
681	Rev	=	46
684	Rev	1X	2
684	Rev	=	95
688	Rev	1X	2
688	Rev	=	9
700	Rev	1X	2
700	Rev	=	14
702	Rev	=	43
711	Rev	1X	3
711	Rev	=	83
728	Rev	1I24=1D	63
728	Rev	1X	3
730	Rev	=	3
752	Rev	1X	8
752	Rev	1X19=1X	2
752	Rev	=	158
753	Rev	1I10=1D31=1X	2
753	Rev	=	35
756	Rev	=	25
778	Rev	1X	50
778	Rev	=	104
798	Rev	1X	4
821	Rev	1X	3
821	Rev	=	106
828	Rev	3D30=3D	61
828	Rev	=	47
830	Rev	=	139
839	Rev	=	91
840	Rev	1X	7
846	Rev	=	44
851	Rev	1X	1
851	Rev	=	29
853	Rev	=	5
856	Rev	=	40
869	Rev	3D	15
871	Rev	1X	6
884	Rev	1X	12
884	Rev	=	53
887	Rev	=	13
895	Rev	=	8
902	Rev	=	8
905	Rev	1I32=1D	71
906	Rev	1I6=1D	47
907	Rev	=	18
914	Rev	1X	3
914	Rev	=	32
916	Rev	1X	1
916	Rev	=	23
942	Rev	1X	42
958	Rev	1X	2
958	Rev	=	82
961	Rev	1I15=1D39=1X	17
961	Rev	=	52
963	Rev	1X	78
963	Rev	1X5=1I8=1D	51
963	Rev	=	58
964	Rev	1X	22
970	Rev	1X	2
970	Rev	1X4=1X	16
970	Rev	=	274
977	Rev	1X	21
995	Rev	1X	1
995	Rev	=	57
998	Rev	1X	1
1005	Rev	=	18
1020	Rev	=	13

Table 13: Summary of sanger sequencing of unique screened colonies

[GAGAIP]24:
NNNNNNNNNNNNNNNNNNNNCTAGAATAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGCCGGTGCAATCCCAGGGGCTGAAGCTATTCCAG
GAGCGGGAGCCATACCCGGAGCAGGCGCAATACCAGGTGCCGGTGCAATCCCAGGGGCTGAAGCTATTCCAGGAGCGGGAGCCATACCCGGAGCAGGCGCAATACCAGGTGCCG
GTGCAATCCCAGGGGCTGAAGCTATTCCAGGAGCGGGAGCCATACCCGGAGCAGGCGCAATACCAGGTGCCGGTGCAATCCCAGGGGCTGAAGCTATTCCAGGAGCGGGAGCCAT
ACCCGGAGCAGGCGCAATACCAGGTGCCGGTGCAATCCCAGGGGCTGAAGCTATTCCAGGAGCGGGAGCCATACCCGGAGCAGGCGCAATACCAGGTGCCGGTGCAATCCCAGG
GGCTGAAGCTATTCCAGGAGCGGGAGCCATACCCGGAGCAGGCGCAATACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGAT
CCGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGC
TGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGANGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATT
AAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGNCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTNNNCCTTCCTTTCTCGCCNCGTTCGCCGGCTTTCCCCGTC
AAGCTCTAAATCGGGGGCTCCCTTTAGGGNTCCNATTTAGTGCTTTACNGGCACCTCGACCCCAAAAACNTGATTANNGTGATGGNNNNGTNNNGGCNTCNNCCNGANNNANNN
GNNTTTCGCCNTTGNCNTNGANTCNCNTNNTANNNNGGNNNNNNNNNAACNNNNNNNNNNNNNNNCNCNGNNNNNTTNNNTTNNANGGNNNTNNGCNNATTNNNCNNN
NTNNNNNNNAA

[GAGAIPGAGAIPGAGAIPVAGAIP]6:
NNNNNNNNNNNNNNNNNTNNNNNCTANNNNANTTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGCCGGTGCAATCCCAGGGGCTGGAGCTATTCC
AGGAGCGGGAGCCATACCCGTAGCAGGCGCAATACCAGGTGCCGGTGCAATCCCAGGGGCTGGAGCTATTCCAGGAGCGGGAGCCATACCCGTAGCAGGCGCAATACCAGGTGC
CGGTGCAATCCCAGGGGCTGGAGCTATTCCAGGAGCGGGAGCCATACCCGTAGCAGGCGCAATACCAGGTGCCGGTGCAATCCCAGGGGCTGGAGCTATTCCAGGAGCGGGAGCC
ATACCCGTAGCAGGCGCAATACCAGGTGCCGGTGCAATCCCAGGGGCTGGAGCTATTCCAGGAGCGGGAGCCATACCCGTAGCAGGCGCAATACCAGGTGCCGGTGCAATCCCAG
GGGCTGGAGCTATTCCAGGAGCGGGAGCCATACCCGTAGCAGGCGCAATACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGA
TCCGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCG
CTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGACTATATCCGGATTGNCGAATGGGACGCGCCCTGTAGCGGCGCATTT
AAGCGCGGCGGGNGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTC
AAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCNNATTTANNGCTTNNNCNNNCGACCCCAAAAACTTGATNNNGTGANGNNCACGTANTGGGCCATCGCCNGNNANANNTTTN
NNNCCNTTNNACNTTNNN

[GANMPPGASIPPGANIPPGASIPP]7:
NNNNNNNNNNNGNNNNNTCNNNCTNNNNNAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCTGGTGCCAACATGCCTCCAGGAGCGAGCATACCGCCCG
GGGCTAACATTCCTCCGGGTGCAAGCATCCCACCTGGTGCCAACATGCCTCCTGGAGCGAGCATACCGCCCGGGGCTAACATTCCTCCGGGTGCAAGCATCCCACCTGGTGCCAAC
ATGCCTCCAGGAGCGAGCATACCGCCCGGGGCTAACATTCCTCCGGGTGCANNCATCCCACCTGGTGCCAACATGCCTCCAGGAGCGAGCATACCGCCCGGGGCTAACATTCCTCC
GGGTGCAAGCATCCCACCTGGTGCCAACATGCCTCCAGGAGCGAGCATACCGCCAGGGGCTAACATTCCTCCGGGTGCAAGCATCCCACCTGGTGCCAACATGACTCCAGGAGCG
AGCATACCGCCAGGGNCTAACATTCCTCCGGGTGCAAGCATCCCACCTGGNGCCAACATGCCTCCAGGAGCGAGCATACCGCCAGGGGCTAACATTCCTCCGGGTGCAAGCATCC
CACCTGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCNNGATCCNTATTCNAGCTCCATCGACAAGCTTGNNGNCGCACTCNANCACCA
NCNCCACCACCACTGANATCCNGCTGCTAACAAAGCCCGAAAGGAAGCTGANTTGGCTGCTGCCANCGNTGATCAATAACTAGCATANNNGNNTGNGGCCTCTAAACGGGGTCT
TGAGGGGTTTTTTGCTGAGAGGANGAACTATATCCGGANTGNNNAATGGNNNGCGCCCTGNAGCNGCGCATTAAGCGNNNCNNGCGTGNNTGTNTACGCGCANNNNGANNNC
NNCNCNTGCCAGCNNCCCTANCGCCNGCTCCTTNNCNNTNNNCNNNTACNNTNNTCGNCANNATCNCCAGCNTNTCNNNNGNTNAGCTACTNNNNNNGGNNNCTNCCNTT
ANGGNTTCCGANTNGANNGNNNNNNGNTNCCTCNNNNNNNNNNAANNNTNNNNNNNNNNNNNNNNNNNNNCTNGNCCATNNNCNNNGACNNNANNNNNNNNNNNC
NNNGNCNNTNNNNCNANNNNCNTNNNNNNNNNNNNNNNTCCANCTNGANANCANN

[GANMPPGASIPPGANIPPGASIPP]8:
NNNNNNNNNNNNNNTNNNNNCTAGAATAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCTGGTGCCAACATGCCTCCAGGAGCGAGCATACCGCCAGGG
GCTAACATTCCTCCGGGTGCAAGCATCCCACCTGGTGCCAACATGCCTCCAGGAGCGAGCATACCGCCAGGGGCTAACATTCCTCCGGGTACAAGCATCCCACCTGGTGCCAACAT
GCCTCCAGGAGCGAGCATACCGCCAGGGGCTAACATTCCTCCGGGTGCAAGCATCCCACCTGGTGCCAACATGCCTCCAGGAGCGAGCATACCGCCAGGGGCTAACATTCCTCCGG
GTGCAAGCATCCCACCTGGTGCCAACATGCCTCCAGGAGCGAGCATACCGCCAGGGGCTAACATTCCTCCGGGTGCAAGCATCCCACCTGGTGCCAACATGCCTCCAGGAGCGAGC
ATACCGCCAGGGGCTAACATTCCTCCGGGTGCAAGCATCCCACCTGGTGCCAACATGCCTCCAGGAGCGAGCATACCGCCAGGGGCTAACATTCCTCCGGGTGCAAGCATCCCACC
TGGTGCCAACATGCCTCCAGGAGCGAGCATACCGCCAGGGGCTAACATTCCTCCGGGTGCAAGCATCCCACCTGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACG
GNTAANGATGATCTTCAGGATCCGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAG
CTGAGTTGGCTGCTGCCACCGCTGANCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAANGAGNAACTATATCCGGATTNNGAATGGGA
CGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACNCGCANCGNGACCGCTANACTTGCCAGCGGCCCTAGCGCCCGCTCCTTTNNCTTTCTCCCTTNN

[GANMPQGASIPPGANIPPGASIPP]6:
NNNNNNNNNNNNNNNTCNNCTCTANNNNNATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCTGGTGCCAACATGCCTCAAGGAGCGAGCATACCGCCAGG
GGCTAACATTCCTCCGGGTGCAAGCATCCCACCTGGTGCCAACATGCCTCAAGGAGCGAGCATACCGCCAGGGGCTAACATTCCTCCGGGTGCAAGCATCCCACCTGGTGCCAACA
TGCCTCAAGGAGCGAGCATACCGCCAGGGGCTAACATTCCTCCGGGTGCAAGCATCCCACCTGGTGCCAACATGCCTCAAGGAGCGAGCATACCGCCAGGGGCTAACATTCCTCCG
GGTGCAAGCATCCCACCTGGTGCCAACATGCCTCAAGGAGCGAGCATACCGCCAGGGGCTAACATTCCTCCGGGTGCAAGCATCCCACCTGGTGCCAACATGCCTCAAGGAGCGA
GCATACCGCCAGGGGCTAACATTCCTCCGGGTGCAAGCATCCCACCTGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTAT
TCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGANC
AATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCG
CGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCNGCTTTCCCGTCAAGCTC
TAAAATCGGGGGGCTCCCTTTANGGTTCCGATTTANTGCTNACNGCACTCCNACCCCAAAAACTTGNNTAGGGTGANGGNTCANNTNNNGNNCNTCNNCCTGANANANNNNTTT
CNCCNTTGANCNTNNANTCNCNTNNTNAATANNGNCTNNNNNCNANTNNNNCNANANTCNNNNNNNNNNNNNNNNNTNNNTNGATNNNNNN

[GDGYAP]32:
NNNNNNNNNNNNNNNNNNNCCCTCTAGAATAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGATGGTTATGCCCCTGGCGATGGATATGCTCCTG
GAGACGGCTACGCTCCCGGAGATGGGTATGCACCAGGTGATGGTTATGCCCCTGGCGATGGATATGCTCCTGGAGACGGCTACGCTCCCGGAGATGGGTATGCACCAGGTGATGGT
TATGCCCCTGGCGATGGATATGCTCCTGGAGACGGCTACGCTCCCGGAGATGGGTATGCACCAGGTGATGGTTATGCCCCTGGCGATGGATATGCTCCTGGAGACGGCTACGCTCCC
GGAGATGGGTATGCACCAGGTGATGGTTATGCCCCTGGCGATGGATATGCTCCTGGAGACGGCTACGCTCCCGGAGATGGGTATGCACCAGGTGATGGTTATGCCCCTGGCGATGG
ATATGCTCCTGGAGACGGCTACGCTCCCGGAGATGGGTATGCACCAGGTGATGGTTATGCCCCTGGCGATGGATATGCTCCTGGAGACGGCTACGCTCCCGGAGATGGGTATGCAC
CAGGTGATGGTTATGCCCCTGGCGATGGATATGCTCCTGGAGACGGCTACGCTCCCGGAGATGGGTATGCACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACG
GCTAATGATGATCTTCAGGATCCGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGC
TGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTTGGGGCCTCTAAACGGGTCNTGANGGTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGAC
GCGCCCTGTANCGGCNCATTANCGCGGCGGGTGTGNGNTACGCGCAGCGTGACNCTACANTTGCNGCNNNTAGCGCCGCTCTTTCGCTTTCNTNCCNTNCTTTNNNGCNNNNNTC
NCCGNN

[GEGNIP]24:
NNNNNNNNNNNNNNNNNNNNNCTAGAATAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGAGGGTAACATTCCAGGAGAAGGAAATATACCG
GGCGAAGGCAATATCCCAGGGGAAGGTAATATACCAGGTGAGGGTAACATTCCAGGAGAAGGAAATATACCGGGCGAAGGCAATATCCCAGGGGAAGATAATATACCAGGTGAG
GGTAACATTCCAGGAGAAGGAAATATACCGGGCGAAGGCAATATCCCAGGGGAAGGTAATATACCAGGTGAGGGTAACATTCCAGGAGAAGGAAATATACCGGGCGAAGGCAAT
ATCCCAGGGGAAGGTAATATACCAGGTGAGGGTAACATTCCAGGAGAAGGAAATATACCGGGCGAAGGCAATATCCCAGGGGAAGGTAATATACCAGGTGAGGGTAACATTCCA
GGAGAAGGAAATATACCGGGCGAAGGCAATATCCCAGGGGAAGGTAATATACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAG
GATCCGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCAC
CGCTGANCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGC
ATTNAANCGCGGCGGGTGTGNTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTANCGCCCGCTCCTTTCGCTTTCTTCCCTTTCTCGCCACGNTCGCCGGCTTTCC
CCGTCAAGCTNTAAATCGGGGGCNNCCTTTAGGGTTCCGATTANTGCTTTACGGCACCTCNACCCCAAAANNTTGATTNNGNNGANGGTTCANGTANNGGGNNATCNNCCNGNT
AGACGGTTTTCGCCNNNNGANGTTGNAGTCCACNTNCTTNANNNATGNNNTCNTGNTNNN

[GGHNIP]24:
NNNNNNNNNNNNNNNNNNNCTANNNNNATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGGGCATAATATCCCTGGCGGTCACAATATACCCGGAG
GTCATAACATTCCTGGAGGACACAACATACCAGGTGGGCATAATATCCCTGGCGGTCACAATATACCCGGAGGTCATAACATTCCTGGAGGACACAACATACCAGGTGGGCATAAT
ATCCCTGGCGGTCACAATATACCCGGAGGTCATAACATTCCTGGAGGACACAACATACCAGGTGGGCATAATATCCCTGGCGGTCACAATATACCCGGAGGTCATAACATTCCTGG
AGGACACAACATACCAGGTGGGCATAATATCCCTGGCGGTCACAATATACCCGGAGGTCATAACATTCCTGGAGGACACAACATACCAGGTGGGCATAATATCCCTGGCGGTCAC
AATATACCCGGAGGTCATAACATTCCTGGAGGACACAACATACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTC

GAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAA
TAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCG
GCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTANCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCNTCAAGCTCT
AAATCGGGGGCTCCCTTTANGGTTCCGATTTNNNGCTTTACNGCACCTCGACCCCAAAAANCTTGATNNNGGTGATGGNTCACGTNNNGGGCCATCNNCCTGATANACNNNTTTT
CGCCNNNGNCGTTGGAGTCNACGTNNNNTANANTGNNTNNNTNCAAACNGGNNCNNNCNTNNNCCNNNNNNNGTCTNNNNNTTTGANTTNNNNNNNAATTNTGACN

[GGHQQP]24:
NNNNNNNNNNNNNNNNGNNNNNNNNTCNNTCTAGAATAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGGGCACCAACAGCCTGGAGGACAT
CAACAACCCGGCGGTCATCAGCAGCCAGGGGGACACCAGCAACCAGGTGGGCACCAACAGCCTGGAGGACATCAACAACCCGGCGGTCATCAGCAGCCAGGGGGACACCAGCA
ACCAGGTGGGCACCAACAGCCTGGAGGACATCAACAACCCGGCGGTCATCAGCAGCCAGGGGGACACCAGCAACCAGGTGGGCACCAACAGCCTGGAGGACATCAACAACCCG
GCGGTCATCAGCAGCCAGGGGGACACCAGCAACCAGGTGGGCACCAACAGCCTGGAGGACATCAACAACCCGGCGGTCATCAGCAGCCAGGGGGACACCAGCAACCAGGTGGG
CACCAACGGCCTGGAGGACATCAACAACCCGGCGGTCATCAGCAGCCAGGGGGACACCAGCAACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAAT
GATGATCTTCAGGATCCGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTT
GGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCC
CTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCG
CCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACNGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGNTCACGTAGTGGGCCATC
NNCCTGAANAGANNGGNTTTNNNNCCCTTTGA

[GGKHAP]28:
NNNNNNNNNNNTTCCCCTCTANNNNAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGGGAAGCACGCACCCGGAGATAAACACGCTCCAGGAG
GAAAACATGCCCCTGGCGGTAAGCATGCACCAGGTGGGAAGCACGCACCCGGAGGTAAACACGCTCCAGGAGGAAAACATGCCCCTGGCGGTAAGCATGCACCAGGTGGGAAGC
ACGCACCCGGAGGTAAACACGCTCCAGGAGGAAAACATGCCCCTGGCGGTAAGCATGCACCAGGTGGGAAGCACGCACCCGGAGGTAAACACGCTCCAGGAGGAAAACATGCCC
CTGGCGGTAAGCATGCACCAGGTGGGAAGCACGCACCCGGAGGTAAACACGCTCCAGGAGGAAAACATGCCCCTGGCGGTAAGCATGCACCAGGTGGGAAGCACGCACCCGGAG
GTAAACACGCTCCAGGAGGAAAACATGCCCCTGGCGGTAAGCATGCACCAGGTGGGAAGCACGCACCCGGAGGTAAACACGCTCCAGGAGGAAAACATGCCCCTGGCGGTAAGC
ATGCACCANGTGAAAACCTGTATTTTCANGGCCATCACCATCACCATCACNGCTAATGATGATCTTCANGATCCGTATTCNAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCA
CCACCACCACCACCACTGANATCCGGCTGCTAACNAAGCCCGAANGGAAGCTGANTTGGCTGCTGCCACCGCTGANCAATAACTATCATAACCCCTTGNGGCCTCTAAACGGGTC
TTGAGGGGTTTTTTGCTGAANNGAGGAACTATATCCGGATTNNGAATGNGACGCGCCCTGTANCGGCGCATTAANCGCGGCGGGTGTGGTGNTTACGCNCANCGTGACCGCTACA
CTTGCCAGCGCCCTANCGCCCGCTCNTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGNNTTTCCCCGTCAANCTCTAAATCNGGGGCTCCCTTTN

[GGKHAP]24:
NNNNNNNNNNNNNNNTCNNNCTAGANTAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGGGAAGCACGCACCCGGAGGTAAACACGCTCCAGGA
GGAAAACATGCCCCTGGCGGTAAGCATGCACCAGGTGGGAAGCACGCACCCGGAGGTAAACACGCTCCAGGAGGAAAACATGCCCCTGGCGGTAAGCATGCACCAGGTGGGAAG
CACGCACCCGGAGGTAAACACGCTCCAGGAGGAAAACATGCCCCTGGCGGTAAGCATGCACCAGGTGGGAAGCACGCACCCGGAGGTAAACACGCTCCAGGAGGAAAACATGCC
CCTGGCGGTAAGCATGCACCAGGTGGGAAGCACGCACCCGGAGGTAAACACGCTCCAGGAGGAAAACATGCCCCTGGCGGTAAGCATGCACCAGGTGGGAAGCACGCACCCGGA
GGTAAACACGCTCCAGGAGGAAAACATGCCCCTGGCGGTAAGCATGCACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATC
CGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCT
GANCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTA
AGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCA
AGCTCTAAATCGGGGGCTCCCTTTAGGGNNCNATTTNNNGCTTTACGGCACCTCNACCCCAAAAACTTGANTAGGGTGATGGTTCNNCGTANNNGGGCCATCGCCCNGNANANN
NNN

[PGGKHAPGGKHAPGGKHAPDDKHAP]6:
NNNNNNNNNNNNNNNNTCNNCTCTAGAATAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGGGAAGCACGCACCCGGAGGTAAACACGCTCCA
GGAGGAAAACATGCCCCTGACGATAAGCATGCACCAGGTGGGAAGCACGCACCCGGAGGTAAACACGCTCCAGGAGGAAAACATGCCCCTGACGATAAGCATGCACCAGGTGGG
AAGCACGCACCCGGAGGTAAACACGCTCCAGGAGGAAAACATGCCCCTGACGATAAGCATGCACCAGGTGGGAAGCACGCACCCGGAGGTAAACACGCTCCAGGAGGAAAACA
TGCCCCTGACGATAAGCATGCACCAGGTGGGAAGCACGCACCCGGAGGTAAACACGCTCCAGGAGGAAAACATGCCCCTGACGATAAGCATGCACCAGGTGGGAAGCACGCACC
CGGAGGTAAACACGCTCCAGGAGGAAAACATGCCCCTGACGATAAGCATGCACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAG
GATCCGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCAC
CGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGC
ATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCC
GTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCNATTTAGTGCTTTACGGCACCTCGACCCCAAAAACTTGATTNNNTGANGGTNACGTANNGNNNTCGCCCTGATAGANGNT
TTCGCCTTTGACGTGNNNTCACNTTCTTTANNNTGGNNNNTNNTCCN

[GGLHTP]24:
NNNNNNNNNNNNNTCCNCTCTAGAATAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGGCTTGCATACTCCGGGAGGTTTACACACTCCAGGCG
GTCTACATACCCCAGGAGGACTTCATACACCAGGTGGCTTGCATACTCCGGGAGGTTTACACACTCCAGGCGGTCTACATACCCCAGGAGGACTTCATACACCAGGTGGCTTGCATA
CTCCGGGAGGTTTACACACTCCAGGCGGTCTACATACCCCAGGAGGACTTCATACACCAGGTGGCTTGCATACTCCGGGAGGTTTACACACTCCAGGCGGTCTACATACCCCAGGA
GGACTTCATACACCAGGTGGCTTGCATACTCCGGGAGGTTTACACACTCCAGGCGGTCTACATACCCCAGGAGGACTTCATACACCAGGTGGCTTGCATACTCCGGGAGGTTTACAC
ACTCCAGGCGGTCTACATACCCCAGGAGGACTTCATACACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCGAG
CTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAA
CTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCG
GGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAA
TCGGGGGCTCCCTTTANGGTTCCNATTTAGTGCTTTANNGCACCTCGACCCAAAAACTTGATTNNGGTGANGGTTCACGTANNGGGCCATCNCCCTGATANACGGNTTTTN

[GGLHTP]28:
NNNNNNNNNNNNNNTCCNTCTAGAATAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGGCTTGCATACTCCGGGAGGTTTACACACTCCAGGCGGT
CTACATACCCCAGGAGGACTTCATACACCAGGTGGCTTGCATACTCCGGGAGGTTTACACACTCCAGGCGGTCTACATACCCCAGGAGGACTTCATACACCAGGTGGCTTGCATACT
CCGGGAGGTTTACACACTCCAGGCGGTCTACATACCCCAGGAGGACTTCATACACCAGGTGGCTTGCATACTCCGGGAGGTTTACACACTCCAGGCGGTCTACATACCCCAGGAGG
ACTTCATACACCAGGTGGCTTGCATACTCCGGGAGGTTTACACACTCCAGGCGGTCTACATACCCCAGGAGGACTTCATACACCAGGTGGCTTGCATACTCCGGGAGGTTTACACAC
TCCAGGCGGTCTACATACCCCAGGAGGACTTCATACACCAGGTGGCTTGCATACTCCGGGAGGTTTACACACTCCAGGCGGTCTACATACCCCAGGAGGACTTCATACACCAGGTG
AAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCAC
CACTGAGATCCGGCTGCTAACAAAGCCCGANAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTT
GCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCT
AGCGCCCGCTCCTTTCNCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCANCTCTAANTCNGGGGCTCCTTTAGGGTTCCGATTTAATNNTTTACGNACCNCGANN
CCCANAACNTNATTNNGGNGNNNN

[GGMLAP]28:
NNNNNNNNNNNNNNNNNNNCTANNNNNATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGGGATGCTTGCACCCGGAGGAATGTTAGCCCCTGGC
GGTATTCTCGCTCCTGGAGGCATGTTGCCACCAGGTGGGATGCTTGCACCCGGAGGAATGTTAGCCCCTGGCGGTATGCTCGCTCCTGGAGGCATGTTGCCACCAGGTGGGATGCTT
GCACCCGGAGGAATGTTAGCCCCTGGCGGTATGCTCGCTCCTGGAGGCATGTTGCCACCAGGTGGGATGCTTGCACCCGGAGGAATGTTAGCCCCTGGCGGTATGCTCGCTCCTGGA
GGCATGTTGCCACCAGGTGGGATGCTTGCACCCGGAGGAATGTTAGCCCCTGGCGGTATGCTCGCTCCTGGAGGCATGTTGCCACCAGGTGGGATGCTTGCACCCGGAGGAATGTTA
GCCCCTGGCGGTATGCTCGCTCCTGGAGGCATGTTGCCACCAGGTGGGATGCTTGCACCCGGAGGAATGTTAGCCCCTGGCGGTATGCTCGCTCCTGGAGGCATGTTGCCACCAGGT
GAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCA
CCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTT
GCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTANCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCANCGTGACCGCTACACTTGCCAGCGCCC
TANCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGNNN

[GGMLAP]24:
NNNNNNNNNNNNNNNNNNNTCTAGAATAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGGGATGCTTGCACCCGGAGGAATGTTAGCCCCTGG
CGGTATGCTCGCTCCTGGAGGCATGTTGGCACCAGGTGGGATGCTTGCACCCGGAGGAATGTTAGCCCCTGGCGGTATGCTCGCTCCTGGAGGCATGTTGGCACCAGGTGGGATGCT
TGCACCCGGAGGAATGTTAGCCCCTGGCGGTATGCTCGCTCCTGGAGGCATGTTGGCACCAGGTGGGATGCTTGCACCCGGAGGAATGTTAGCCCCTGGCGGTATGCTCGCTCCTGG
AGGCATGTTGGCACCAGGTGGGATGCTTGCACCCGGAGGAATGTTAGCCCCTGGCGGTATGCTCGCTCCTGGAGGCATGTTGGCACCAGGTGGGATGCTTGCACCCGGAGGAATGT
TAGCCCCTGGCGGTATGCTCGCTCCTGGAGGCATGTTGGCACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCGA
GCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATA
ACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCNGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGC

GGGTGTGGTGGTTACGCGCANCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCNTCCCTTCCTTTCTCGCACGTCGCNGCTTTCCCCGTCAAGCTCTAATCG
GGCTCCNTTNGGNTCCGATTANTGCTTNCGNNCTCGACCCAAAACTTGATAGGNGANGGTTCACGTNNGGGCNNTNNNCCNNN

[GGMLAPGGMLAPGGMLAPGGMFAP]6:
NNNNNNNNGNNNNNTCCNTCTAGANNNATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGGGATGCTTGCACCCGGAGGAATGTTAGCCCCTGGCG
GTATGCTCGCTCCTGGAGGCATGTTTGCACCAGGTGGGATGCTTGCACCCGGAGGAATGTTAGCCCCTGGCGGTATGCTCGCTCCTGGAGGCATGTTTGCACCAGGTGGGATGCTTG
CACCCGGAGGAATGTTAGCCCCTGGCGGTATGCTCGCTCCTGGAGGCATGTTTGCACCAGGTGGGATGCTTGCACCCGGAGGAATGTTAGCCCCTGGCTGTATGCTCGCTCCTGGAG
GCATGTTTGCACCAGGTGGGATGCTTGCACCCGGAGGAATGTTAGCCCCTGGCGGTATGCTCGCTCCTGGAGGCATGTTTGCACCAGGTGGGATGCTTGCACCCGGAGGAATGTTAG
CCCCTGGCGGTATGCTCGCTCCTGGAGGCATGTTTGCACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCGAGCT
CCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCNGCCACCGCTGANCAATAACTAGC
AGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGG
TGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCG
GGGGCTCCCTTTACGGTTCCNATTTANTGCTTTANNNACCTCGACCCCAAAAACTTGATTNNNTGANGGNTCACGNN

[GGQHIP]24:
NNNNNNNNNNNNNNNTTNCCNCTAGAATAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGGGCAGCACATTCCTGGCGGTCAACATATACCCGG
AGGTCAGCATATTCCAGGGGGACAACACATACCAGGTGGGCAGCACATTCCTGGCGGTCAACATATACCCGGAGGTCAGCATATTCCAGGGGGACAACACATACCAGGTGGGCAG
CACATTCCTGGCGGTCAACATATACCCGGAGGTCAGCATATTCCAGGGGGACAACACATACCAGGTGGGCAGCACATTCCTGGCGGTCAACATATACCCGGAGGTCAGCATATTCC
AGGGGGACAACACATACCAGGTGGGCAGCACATTCCTGGCGGTCAACATATACCCGGAGGTCAGCATATTCCAGGGGGACAACACATACCAGGTGGGCANCACATTCCTGGCGGT
CAACATATACCCGGAGGTCAGCATATTCCAGGGGGACAACACATACCAGGTGAAAACCTGTATTTTCACGGCCATCACCATCACCATCACGGGTAATGATGATCTTCAGGATCCGT
ATTCGAGCTCCTTCAACAAGCTTGCGGCCGCACTCNANCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCNGCCACCGCTGA
NCAATAACTAGCATAACCCCTTGGGGCNTCTAAACGNGTCTTGANGGGTTTTTTTGCTGAAAGGAGGANCTATATCCGGATTGCCGAATGGGACGCGCCCTGTAGCGGCCACATTT
AAGCGCGCCCGGGTGTGGTGGTTACNCGCAACCGAGACATTTACACTTTNCAGGCGCCCNAACNCCCGCCTCCTTATCGCTTTCCTTCCCATCCTTTTNCGNCACGTTCNNCGNGNT
TTCCCANNCAANNTANAAAACNGGNNGCNCCCTTNTANGGTNTCNGAGTTNNAGGTTTACGGCANCANNACCCCAANGAACTTNNNATNNNNGAGNGNNNCACTCNNGGCN
NAATCNCCCCTANANGACGNGTTTNCACCCCTTTGCACTTGAAATNNNNGTGATTATAACGANNNGAGNNTTTGTCNCNACNAGCANNNNNNCACTGNCCNGTNCGNNNGCN
NNNTNGNNNANNNNGANAAGNNN

[GYGYAP]24:
NNNNNNNNNNNNNNNNNNNNNCTNNNNNNATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTTATGGGTACGCTCCTGGATACGGATATGCCCCAGG
GTATGGCTATGCTCCCGGTTACGGCTACGCACCAGGTTATGGGTACGCTCCTGGATACGGATATGCCCCAGGGTATGGCTATGCTCCCGGTTACGGCTACGCACCAGGTTATGGGTA
CGCTCCTGGATACGGATATGCCCCAGGGTATGGCTATGCTCCCGGTTACGGCTACGCACCAGGTTATGGGTACGCTCCTGGATACGGATATGCCCCAGGGTATGGCTATGCTCCCGG
TTACGGCTACGCACCAGGTTATGGGTACGCTCCTGGATACGGATATGCCCCAGGGTATGGCTATGCTCCCGGTTACGGCTACGCACCAGGTTATGGGTACGCTCCTGGATACGGATA
TGCCCCAGGGTATGGCTATGCTCCCGGTTACGGCTACGCACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCGA
GCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGANTTGGCTGCTGCCACCGCTGANCAATA
ACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGNTTTTTGCTGAAAGGAGGAACTATATCCGGATTNGGCGAATGGGACGCGCCCTGTAAGCGGCGCNTTAAGCGCG
GCGGGTGTGGTGGNTTACGCGCAACGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCNCTTCCTTCTNGCCACGTCGCCGGCTTTCCCCNTCAAGCTNT
AAATCNGGGGCTCCCTTTAGGNTCCNATNNNNGCTTACGGNNCTCNACCCCAAAAACTTGANTNNGGNNNANGGNTCANGTAGTGGNCCATNNCCNTNGANANANANGNTTT
NNNGCCCNNN

[GGNNIP]24:
NNNNNNNNNNNNNNNCCCTCTAGAATAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGGCAACAATATACCGGGCGGAAATAATATCCCAGGG
GGGGGTAATAACATTCCAGGAGGAAACAACATACCAGGTGGCAACAATATACCGGGCGGAAATAATATCCCAGGGAGTAATAACATTCCAGGAGGAAACAACATACCAGGTGGC
AACAATATACCGGGCGGAAATAATAACATTCCAGGAGGAAACAACATACCAGGTGGCAACAATATACCGGGCGGAAATAATATCCCAGGGGGTAATAACATTCCAGGAGGAAAC
AACATACCAGGTGGCAACAATATACCGGGCGGAAATAATAACATTCCAGGAGGAAACAACATTCCAGGAGGAAACAACATACCAGGTGGCAACAATATACCGGGCGGAAATAAT
ATCCCAGGGGGTAATAACATTCCAGGAGGAAACAACATACCGGTATATTGTTGCCACCTGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTT
CAGGATCCGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGANATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGC
CACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGANGAACTATATCCCNGATTGGCCGAATGGGGACGCGCCCCTGTA
GCGGCGCANTTAAGCGCGGGCGGGTGTGGNGGGTTACNCGNANNNNGACCGCTACACTTGNNNNNNNCCNTANCGCCCGCCTCCNTTTCGCTTTCNTCNCNTCNTTTCTCNCCA
CNNTCGNCNGCTTTTCCCCNTTNANGCTCTTAAATCNGGGGGNNCCCTTTANGGNTNCNNATTNNNNGCTTTACGGNACCTCGAACCCGNAANAACTTGNATTAGTGGNNNTGN
NTCACGGTNNNGNGCCNNCGNNCNGNATNNNCNGGTNNTTCANNNCN

[GHIGVP]24:
NNNNNNNNNNNNNNNNNNTCTNNNNNNATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTCATATTGGCGTACCCGGTCACATTGGTGTTCCTGGGCA
TATCGGAGTCCCTGGACATATAGGCGTACCAGGTCATATTGGCGTACCCGGTCACATTGGTGTTCCTGGGCATATCGGAGTCCCTGGACATATAGGCGTACCAGGTCATATTGGCGT
ACCCGGTCACATTGGTGTTCCTGGGCATATCGGAGTCCCTGGACATATAGGCGTACCAGGTCATATTGGCGTACCCGGTCACATTGGTGTTCCTGGGCATATCGGAGTCCCTGGACAT
ATAGGCGTACCAGGTCATATTGGCGTACCCGGTCACATTGGTGTTCCTGGGCATATCGGAGTCCCTGGACATATAGGCGTACCAGGTCATATTGGCGTACCCGGTCACATTGGTGTTC
CTGGGCATATCGGAGTCCCTGGACATATAGGCGTACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCGAGCTCC
GTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAG
CATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTG
TGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGG
GGGCTCCCTTTAGGGTTCCGATTTNNTGCTTTANNNNACCTCGACCCCAAAAANTTGATNANGGTGANGGNTCACGNNNNGNNCATCNNCCTGANNNANGGTTTTTCGCCCTTTG
ACGTTGGNGTCCACGTTNNTNANANNGGNNTCNNGNTCCNNNNN

[GHMGTP]24:
NNNNNNNNNNNNNNNNNNNCTAGAATAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTCACATGGGAACTCCGGGTCATATGGGTACTCCAGGG
CATATGGGGACGCCAGGACATATGGGCACACCAGGTCACATGGGAACTCCGGGTCATATGGGTACTCCAGGGCATATGGGGACGCCAGGACATATGGGCACACCAGGTCACATGG
GAACTCCGGGTCATATGGGTACTCCAGGGCATATGGGGACGCCAGGACATATGGGCACACCAGGTCACATGGGAACTCCGGGTCATATGGGTACTCCAGGGCATATGGGGACGCC
AGGACATATGGGCACACCAGGTCACATGGGAACTCCGGGTCATATGGGTACTCCAGGGCATATGGGGACGCCAGGACATATGGGCACACCAGGTCACATGGGAACTCCGGGTCAT
ATGGGTACTCCAGGGCATATGGGGACGCCAGGACATATGGGCACACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGT
ATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGA
NCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGNCGAATGGGACGCGCCCTGTAGCGGCGCATTAAG
CGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAG
CTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTNNCNGCACNTCGACCCCAAAAAACTTGATTANGGNGATGGNTCACGTANTGNNCATCGCCCTGATAGACGNTTT
TCGCCNTTTGACGTTGNNTCNNCGTTNNNNNANTGGANNNNNGTCCAACNGNNNCAANNCTCANNNNANNNCNGNCTNTTCNTTTGATNNNNNNNNANNNN

[GYGHAP]28:
NNNNNNNNNNNNNTNNNNCTNNNNNNATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTTATGGACATGCTCCGGGGTATGGTCATGTGCCTGGCTA
CGGTCACGCTCCTGGATATGGGCATGCACCAGGTTATGGACATGCTCCGGGGTATGGTCATGCGCCTGGCTACGGTCACGCTCCTGGATATGGGCATGCACCAGGTTATGGACATGC
TCCGGGGTATGGTCATGCGCCTGGCTACGGTCACGCTCCTGGATATGGGCATGCACCAGGTTATGGACATGCTCCGGGGTATGGTCATGCACCTGGCTACGGACACGCTCCTGGATA
TGGGCATGCACCAGGTTATGGACATGCTCCGGGGTATGGTCATGCGCCTGGCTACGGTCACCCTCCTGGATATGGGCATGCACCACGTTATGGACATGCTCCGGGGTATGGTCATGC
GCCTGGCTACGGTCACGCTCCTGGATATGGGCATGCACCATGTTATGGACATGCTCCGGGGTATGGTCATGCGCCTGGCTACGGTCACGCTCCTGGATATGGGCATGCACCAGGTGA
AAACCTGTATTTTCAGGGCCATCATCACCATCACNGCTAATGATGATCTTCTNGATCCGTATNCNAGCTCNNTCGACNNGCTTGAGACNGNACTCTANCACNNCCNCNACCA
NNNNTGANATNCGGATGCTANCAAANTNCGAAAGGAAGCTGANTTNNNTGCTGCNACCGCTGANNATTAACTANCANNANNCNCTTGGGTACCTCTAAATNGNGTCTTGAGGG
GNTNTTTTGCTGANANNANGAGCTATAANCCGGATTGGTGAATGGTACGCGNNNNGTANCNGCGCNTTANNCGCNGCGNGAAGTGGTGGTTACNNNNAACGNGACAGCTANN
NNNGNNANCNNNNTANCGTCCNGCTCCNTANNATNTCTNCNATNCNTTTCNNNN

[GNGNVP]24:
NNNNNNNNNNNNNNNNNNNNNCTANNNNAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTAATGGTAACGTCCCTGGCAACGGCAATGTTCCA
GGAAATGGAAACGTACCGGGGAATGGGAACGTACCAGGTAATGGTAACGTCCCTGGCAACGGCAATGTTCCAGGAAATGGAAACGTACCGGGGAATGGGAACGTACCAGGTAAT
GGTAACGTCCCTGGCAACGGCAATGTTCCAGGAAATGGAAACGTACCGGGGAATGGGAACGTACCAGGTAATGGTAACGTCCCTGGCAACGGCAATGTTCCAGGAAATGGAAACG
TACCGGGGAATGGGAACGTACCAGGTAATGGTAACGTCCCTGGCAACGGCAATGTTCCAGGAAATGGAAACGTACCGGGGGGGAATGGGAACGTACCAGGTAATGGTAACGTCCC
TGGCAACGGCAATGTTCCAGGAAATGGAAACGTACCGGGGAATGGGAACGTACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAG
GATCCGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCAC
CGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGC
ATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCC

GTCAAGCTCTAAATCGGGGGCTCCCTTTNAGGGTTCCGATTTAGNGCTTTACGGCACCTCGACCCCAAAAAACTTGGATTANGGTGATGGTTTNNCGTANTGGGCCATCNNCCTGAT
AGANGTTTTCGCCCNTTGACNNNNNTCACGTTCTTNAANAGNGGNNNNTNTNCNANNNGGNNNANNNNNTCANNN

[GLSGSP]28:
NNNNNNNNNNNNNNNNNNNNNNNCTANNNNAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTTTATCAGGCAGCCCAGGGTTGAGCGGATCTCC
AGGATTAAGTGGTAGTCCCGGTCTTTCTGGCTCACCAGGTTTATCAGGCAGCCCAGGGTTGAGCGGATCTCCAGGATTAAGTGGTAGTCCCGGTCTTTCTGGCTCACCAGGTTTATCA
GGCAGCCCAGGGTTGAGCGGATCTCCAGGATTAAGTGGTAGTCCCGGTTTTTCTGGCTCACCAGGTTTATCAGGCAGCCCAGGGTTGAGCGGATCTCCAGGATTAAGTGGTAGTCCC
GGTCTTTCTGGCTCACCAGGTTTATCAGGCAGCCCAGGGTTGAGCGGATCTCCAGGATTAAGTGGTAGTCCCGGTCTTTCTGGCTCACCAGGTTTATCAGGCAGCCCAGGGTTGAGCG
GATCTCCAGGATTAAGTGGTAGTCCCGGTCTTTCTGGCTCACCAGGTTTATCAGGCAGCCCAGGGTTGAGCGGATCTCCAGGATTAAGTGGTAGTCCCGGTCTTTCTGGCTCACCAGG
TGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCAGGATCGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACC
ACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATANNCCCTTGGGGCCTCTAAACGGGTCTTGANGGGTTTT
TTGCTGAAAGGAGGAACTATATCCGGATTGNCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGNGNTTACGCGCANCGNGANCNGCTACACTTGCCAGCG
CCCTANCGCCCGCTCCTTTCGCTTTCTTCCCTNCNTTTCTCGCCACNNTCGCNNGNTTTCCCGTCAAGCTCTAATTCGGGGGCTNCCNTTNAGGGTTCCGATTTANNNGCTTTANNGN
ANCNTNNACCCCAAAAACTGATTANGGTGATNNNNNNNTNNGGNNGTCGCCNNGANNNACGNNTTTCGCCCNTNGACNNNGGAAGTCCACGTTCNTNNANNNN

[GSMGSP]24:
NNNNNNNNNNNNNNNNNNNNNNCTAGAATAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTAGTATGGGTTCGCCTGGATCAATGGGCAGTCCT
GGCTCGATGGGATCTCCCGGAAGCATGGGGTCACCAGGTAGTATGGGTTCGCCTGGATCAATGGGCAGTCCTGGCTCGATGGGATCTCCCGGAAGCATGGGGTCACCAGGTAGTAT
GGGTTCGCCTGGATCAATGGGCAGTCCTGGCTCGATGGGATCTCCCGGAAGCATGGGGTCACCAGGTAGTATGGGTTCGCCTGGATCAATGGGCAGTCCTGGCTCGATGGGATCTCC
CGGAAGCATGGGGTCACCAGGTAGTATGGGTTCGCCTGGATCAATGGGCAGTCCTGGCTCGATGGGATCTCCCGGAAGCATGGGGTCACCAGGTAGTATGGGTTCGCCTGGATCAA
TGGGCAGTCCTGGCTCGATGGGATCTCCCGGAAGCATGGGGTCACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTAT
TCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGANTTGGCTGCTGCCACCGCTGANC
AATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCG
CGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTANCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCT
CTAAATCGGGGGCTNCCTTTANGGTTCCNATTTNNGCTTTACGGNNCCTCGACNCNAAAACTTGATNNNGTGATGGNTCACGTANGGGNCATCNCCCTGANANANGGTTTTCNNN

[GIGQAP]28:
NNNNNNNNNNNNNATTCCNNCTNNNNNNATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTATCGGTCAAGCGCCAGGAATAGGTCAGGCTCCGGG
TATAGGACAAGCCCCTGGGATTGGACAGGCACCAGGTATCGGTCAAGCGCCAGGAATAGGTCAGGCTCCGGGTATAGGACAAGCCCCTGGGATTGGACAGGCACCAGGTATCGGT
CAAGCGCCAGGAATAGGTCAGGCTCCGGGTATAGGACAAGCCCCTGGGATTGGACAGGCACCAGGTATCGGTCAAGCGCCAGGAATAGGTCAGGCTCCGGGTATAGGACAAGCC
CCTGGGATTGGACAGGCACCAGGTATCGGTCAAGCGCCAGGAATAGGTCAGGCTCCGGGTATAGGACAAGCCCCTGGGATTGGACAGGCACCAGGTATCGGTCAAGCGCCAGGAA
TAGGTCAGGCTCCGGGTATAGGACAAGCCCCTGGGATTGGACAGGCACCAGGTATCGGTCAAGCGCCAGGAATAGGTCAGGCTCCGGGTATAGGACAAGCCCCTGGGATTGGACA
GGCACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACC
ACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGANCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTG
AGGGGTTTTTTGCTGAAAGGANGAACTATATCCGGATTGGCGAATGGGANGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGNGTGGTGGTTACNNNCANCGTGACCGCTANNCT
TGCCAGCGNCCTAGCNNCCGCTCNTTCGNTTTCTTCCCTTCCTTTNNNGCCACGTTCGNNGGCTTTCCCGTCAAGCTCTAAATNNNNN

[GMPIPPGIPIPPEYPSLPVSPMPP]6:
NNNNNNNNNNNNNTCCNTCTANNNNNATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCTGGTATGCCCATTCCGCCTGGGATTCCCATACCTCCGGAATACC
CATCCCTCCCGGTATCTCCCATGCCACCTGGTATGCCCATTCCGCCTGGGATTCCCATACCTCCGGAATACCCATCCCTCCCGGTATCTCCCATGCCACCTGGTATGCCCATTCCGCCT
GGGATTCCCATACCTCCGGAATACCCATCCCTCCCGGTATCTCCCATGCCACCTGGTATGCCCATTCCGCCTGGGATTCCCATACCTCCGGAATACCCATCCCTCCCGGTATCTCCCA
TGCCACCTGGTATGCCCATTCCGCCTGGGATTCCCATACCTCCGGAATACCCATCCCTCCCGGTATCTCCCATGCCACCTGGTATGCCCATTCCGCCTGGGATTCCCATACCTCCGGA
ATACCCATCCCTCCCGGTATCTCCCATGCCACCTGGTGAAAACCTGTATTTTCANGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCGAGCTCCGTCGA
CAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAA
CCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTG
GTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTANCGCCCGCTCCTTTCGCTTTCNTCCCTTCCTTTCTCGCCACGNTCGCCGGCTTTCCCCGTCAAGCTCTAAATCNGGGCTC
CCTTTAGGGNTCCNATTTANTGCTTTACNGCACCTCGACCCCNAAAACTTGATAN

[GQSGLP]24:
NNNNNNNNNNNNATTCCNTCTAGAATAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTCAAAGTGGACTCCCAGGACAATCTGGCTTACCCGGAC
AGAGCGGTCTTCCAGGGCAATCAGGCTTACCAGGTCAAAGTGGACTCCCAGGACAATCTGGCTTACCCGGACAGAGCGGTCTTCCAGGGCAATCAGGCTTACCAGGTCAAAGTGGA
CTCCCAGGACAATCTGGCTTACCCGGACAGAGCGGTCTTCCAGGGCAATCAGGCTTACCAGGTCAAAGTGGACTCCCAGGACAATCTGGCTTACCCGGACAGAGCGGTCTTCCAGG
GCAATCAGGCTTACCAGGTCAAAGTGGACTCCCAGGACAATCTGGCTTACCCGGACAGAGCGGTCTTCCAGGGCAATCAGGCTTACCAGGTCAAAGTGGACTCCCAGGACAATCTG
GCTTACCCGGACAGAGCGGTCTTCCAGGGCAATCAGGCTTACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCG
AGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAAT
AACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGG
CGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTA
ANNCGGGGGCTCCCTTTANGGTTCCGATTTANTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCNNN

[GYVGTP]24:
NNNNNNNNNNNNNNTNNNCTANNNNNNTTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTTATGTGGGCACGCCTGGCTATGTCGGTACTCCCGGATAT
GTAGGAACCCCTGGATACGTTGGTACACCAGGTTATGTGGGCACGCCTGGCTATGTCGGTACTCCCGGATATGTAGGAACCCCTGGATACGTTGGTACACCAGGTTATGTGGGCACG
CCTGGCTATGTCGGTACTCCCGGATATGTAGGAACCCCTGGATACGTTGGTACACCAGGTTATGTGGGCACGCCTGGCTATGTCGGTACTCCCGGATATGTAGGAACCCCTGGATAC
GTTGGTACACCAGGTTATGTGGGCACGCCTGGCTATGTCGGTACTCCCGGATATGTAGGAACCCCTGGATACGTTGGTACACCAGGTTATGTGGGCACGCCTGGCTATGTCGGTACTC
CCGGATATGTAGGAACCCCTGGATACGTTGGTACACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCGAGCTCC
GTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGANCAATAACTAG
CATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTG
TGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGG
GGGCTCCCTTTAGGGTTCCNATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGNCATCGCCCNTNNATNNAC

[GTFGTP]24:
NNNNNNNNNNNNNNTCCCTCTANNNNNATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTACATTCGGAACTCCTGGCACATTTGGGACCCCTGGAA
CCTTTGGTACGCCCGGTACTTTTGGAACACCAGGTACATTCGGAACTCCTGGCACATTTGGGACCCCTGGAACCTTTGGTACGCCCGGTACTTTTGGAACACCAGGTACATTCGGAAC
TCCTGGCACATTTGGGACCCCTGGAACCTTTGGTACGCCCGGTACTTTTGGAACACCAGGTACATTCGGAACTCCTGGCACATTTGGGACCCCTGGAACCTTTGGTACGCCCGGTACT
TTTGGAACACCAGGTACATTCGGAACTCCTGGCACATTTGGGACCCCTGGAACCTTTGGTACGCCCGGTACTTTTGGAACACCAGGTACATTCGGAACTCCTGGCACATTTGGGACC
CCTGGAACCTTTGGTACGCCCGGTACTTTTGGAACACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCGAGCTCC
GTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAG
CATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTG
TGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGG
GGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTNCGGCACCTCNACCCNNAAAAACTTGAATTAGGGTGATGGTTCACGTAGNNGGCATNNCCCTGANANACNN

[GTHGTP]24:
NNNNNNNNNNNNATNNNNNCTAGAATAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTACACATGGCACTCCAGGAACTCACGGTACTCCGGGTAC
GCATGGAACCCCTGGGACTCATGGTACACCAGGTACACATGGCACTCCAGGAACTCACGGTACTCCGGGTACGCATGGAACCCCTGGGACTCATGGTACACCAGGTACACATGGC
ACTCCAGGAACTCACGGTACTCCGGGTACGCATGGAACCCCTGGGACTCATGGTACACCAGGTACACATGGCACTCCAGGAACTCACGGTACTCCGGGTACGCATGGAACCCCTGG
GACTCATGGTACACCAGGTACACATGGCACTCCAGGAACTCACGGTACTCCGGGTACGCATGGAACCCCTGGGACTCATGGTACACCAGGTACACATGGCACTCCAGGAACTCACG
GTACTCCGGGTACGCATGGAACCCCTGGGACTCATGGTACACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTANTTC
GAGCTNCNTCGANNAGCTTGCGGCCGCACTCNAGCACCACCNCCNCCACCNCTGANATCCGGCTGCTANNNAAGCCCGAAAGGAAGCTNAATTGGCTGCTGCCACCGCTGAACA
ATAACTAGCATAACCCCCNTGGGGGCCTCTAAACGGTCTTGANGGGNTTTTTGNNNNAAGGAGNAACTATATCCNGATTGGCGAATGGNNNNCNNCCNGNANCGNNCNTTNNC
CNCGGCGGGNGNGGNGGTNANNNNCAACGTGNCCGNTNANCNTNNANNGCCCNNGCGCCCGCTCCNTNNNCTTTNNTNCCNTNCTTTNNNGCCACNNTNNCNNGNTTTNNCC
NNNANNNNTNANNCNNGGGNTCCCCTTNNNNNNNCNANTTNNNNNN

[GHEGSP]28:
NNNNNNNNNNNNNNNNNNNNNCTAGAATAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTCACGAAGGATCGCCTGGGCATGAAGGTAGTCCTGG
ACACGAGGGTTCTCCCGGACATGAGGGGTCACCAGGTCACGAAGGATCGCCTGGGCATGAAGGTAGTCCTGGACACGAGGGTTCTCCCGGACATGAGGGGTCACCAGGTCACGAA
GGATCGCCTGGGCATGAAGGTAGTCCTGGACACGAGGGTTCTCCCGGACATGAGGGGTCACCAGGTCACGAAGGATCGCCTGGGCATGAAGGTAGTCCTGGACACGAGGGTTCTC
CCGGACATGAGGGGTCACCAGGTCACGAAGGATCGCCTGGGCATGAAGGTAGTCCTGGACACGAGGGTTCTCCCGGACATGAGGGGTCACCAGGTCACGAAGGATCGCCTGGGCA

TGAAGGTAGTCCTGGACACGAGGGTTCTCCCGGACATGAGGGGTCACCAGGTCACGAAGGATCGCCTGGGCATGAAGGTAGTCCTGGACACGAGGGTTCTCCCGGACATGAGGGG
TCACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCA
CCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGNCTCTAAACGGGTCTTGA
GGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTG
CCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTCGCCGGCTTTCCCGTCAAGCTCTNAATCGGGGGCTNCCTTAGGGTCCGATTANTGCTTTANNNNC
NCGACCCAAAACTTGATNAGNTGNATGGTNNACNTANTGNN

[GHHGTP]28:
NNNNNNNNNNNNNNTNNNCTANNNNAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTCACCACGGTACTCCGGGTCATCATGGGACTCCAGGGC
ACCATGGAACGCCAGGACATCACGGCACACCAGGTCACCACGGTACTCCGGGTCATCATGGGACTCCAGGGCACCATGGAACGCCAGGACATCACGGCACACCAGGTCACCACG
GTACTCCGGGTCATCATGGGACTCCAGGGCACCATGGAACGCCAGGACATCACGGCACACCAGGTCACCACGGTACTCCGGGTCATCATGGGACTCCAGGGCACCATGGAACGCC
AGGACATCACGGCACACCAGGTCACCACGGTACTCCGGGTCATCATGGGACTCCAGGGCACCATGGAACGCCAGGACATCACGGCACACCAGGTCACCACGGTACTCCGGGTCAT
CATGGGACTCCAGGGCACCATGGAACGCCAGGACATCACGGCACACCAGGTCACCACGGTACTCCGGGTCATCATGGGACTCCAGGGCACCATGGAACGCCAGGACATCACGGCA
CACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCAC
CACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGANCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAG
GGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGC
CAGCGCCCTANCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCNATTTAGTGCTTTACNG
CACCTCNAC

[GNMGSP]28:
NNNNNNNNNNNNNNNTCCCNTCTANNNNNAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTTATNNGGGCTCGCCAGGCAATATGGGTTCTCCGG
GTAATATGGGAAGTCCAGGAAATATGGGGTCACCAGGTAACATGGGCTCGCCAGGCAATATGGGTTCTCCGGGTAATATGGGAAGTCCAGGAAATATGGGGTCACCAGGTAACAT
GGGCTCGCCAGGCAATATGGGTTCTCCGGGTAATATGGGAAGTCCAGGAAATATGGGGTCACCAGGTAACATGGGCTCGCCAGGCAATATGGGTTCTCCGGGTAATATGGGAAGTC
CAGGAAATATGGGGTCACCAGGTAACATGGGCTCGCCAGGCAATATGGGTTCTCCGGGTAATATGGGAAGTCCAGGAAATATGGGGTCACCAGGTAACATGGGCTCGCCAGGCAA
TATGGGTTCTCCGGGTAATATGGGAAGTCCAGGAAATATGGGGTCACCAGGTAACATGGGCTCGCCAGGCAATATGGGTTCTCCGGGTAATATGGGAAGTCCAGGAAATNTGGGGT
CACCAGGNGNNAACCTGTATTTTCAGGGCCATCNCATCACCATCACGGCTAATGATGATCTTTAGGATCCNTATTCGAGCTCCNTCGGATAAAGCTTGCGGNCGCACTCGAGNTAC
CACCCNACCACCNCTGAAAATCCGGCTGCTNACAAACCCTAAAGGAAGNANANTCAGTCTGCTGCCNCCCTGANNNANACTGNNANCCCCTTGNGNNCNCNNAACGNGNNCT
NANNGGGTTNTGTNNTGNAANGANCAANTNNATNCNGAATNNNCNAATNGGTANNNNNNNNTGTATNGGCTCNNNNNNNNNNGTNGGGNGNNNTGNTAACGANNANNCN
NNACANGTANANNTTGCCANNNGCCCNANCGACANNNTCTTNNGNNNTNCTTNCNTNCCNTTNNTNCNGNTNNNNNNNNNCNCANNNAANNNNNTGNNNTNAGGGANCG

[GGHQAP]28:
NNNNNNNNNNNNNNNNNNCTAGAATAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGGGCATCAGGCTCCTGGCGGTCATCAAGCCCCTGGAG
GTCACCAAGCTCCCGGAGGACATCAAGCACCAGGTGGGCATCAGGCTCCTGGCGGTCATCAAGCCCCTGGAGGTCACCAAGCTCCCGGAGGACATCAAGCACCAGGTGGGCATCA
GGCTCCTGGCGGTCATCAAGCCCCTGGAGGTCACCAAGCTCCCGGAGGACATCAAGCACCAGGTGGGCATCAGGCTCCTGGCGGTCATCAAGCCCCTGGAGGTCACCAAGCTCCCG
GAGGACATCAAGCACCAGGTGGGCATCAGGCTCCTGGCGGTCATCAAGCCCCTGGAGGTCACCAAGCTCCCGGAGGACATCAAGCACCAGGTGGGCATCAGGCTCCTGGCGGTCA
TCAAGCCCCTGGAGGTCACCAAGCTCCCGGAGGACATCAAGCACCAGGTGGGCATCAGGCTCCTGGCGGTCATCAAGCCCCTGGAGGTCACCAAGCTCCCGGAGGACATCAAGCA
CCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCA
CCACCACCACCTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCCGCTGANCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGG
GTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACNCGCAGCGTGANCGCTACACTTGCCA
GCGCCCTANCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCNCGTTCGCCGGCTTTCCCCGTCAAGN

[GYMGKP]24:
NNNNNNNNNNNNNNNNNNNCTANNNNTNNTTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTTATATGGGTAAGCCTGGATATATGGGAAAACCCG
GCTATATGGGCAAGCCAGGATACATGGGGAAACCAGGTTATATGGGTAAGCCTGGATATATGGGCAAGCCAGGATACATGGGGAAACCAGGTTATAT
GGGTAAGCCTGGATATATGGGAAAACCCGGCTATATGGGCAAGCCAGGATACATGGGGAAACCAGGTTATATGGGTAAGCCTGGATATATGGGAAAACCCGGCTATATGGGCAAG
CCAGGATACATGGGGAAACCAGGTTATATGGGTAAGCCTGGATATATGGGAAAACCCGGCTATATGGGCAAGCCAGGATACATGGGGAAACCAGGTTATATGGGTAAGCCTGGAT
ATATGGGAAAACCCGGCTATATGGGCAAGCCAGGATACATGGGGAAACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCC
GTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCNGNNGCTAACAAAGCCCGAAAGGAAGCTGANTTGGCTGCTGCCACCGCT
GANCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTANCGGCGCATT
NAGCGCGGGCGGGGTGTGGGTGGGTTACGCGCANCGTGACCGCTACACTTGCCAGCGCCCTAACGCCCNCTCCTTTTNGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTC
CCCGTNANCNCTANNTCGNNGCTCCCTTTNAGGTTCGANTTANTGCNTACAGANCTCGACCCCAAAAACTTGATNAGGTGANGGTCACGTANGGGGCCTCNNNNNNNAAGACGT
TTTCACCNTTTAACTTGGANN

[GTLGQP]24:
NNNNNNNNNNNNNNNNNNCTNNNNNNATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCTGGTTGTCCTACACTCGGACAGCCTGGAACATTACGTCAAC
CCGGTACACTTGGGCAGCCAGGGACTCTAGGACAACCAGGTACGCTCGGACAGCCTGGAACATTATGTCAACCCGGTACACTTGGGCAGCCAGGGACTCTAGGACAACCAGGTAC
GCTCGGACAGCCTGGAACATTAGGTCAACCCGGTACACTTGGGCAGCCAGGGACTCTAGGACAACCAGGTACGCTCGGACAGCCTGGAACATTACGTCAACCCGGTACACTTGGG
CAGCCAGGGACTCTAGGACAACCAGGTACGCTCGGACAGCCTGGAACATTAGGTCAACCCGGTACACTTGGGCAGCCAGGGACTCTAGGACAACCAGGTACGCTCGGACAGCCTG
GAACATTANGTCAACCCGGTACATTTGGGCAGCCAGGGACTCTAGGACAACCANGTGAAAANNNGTATTTTCNNGGCCATCACCATCACCATCACNGCTAATGATGATCTTCANG
ATCCGTATTCCAGCTCCGTCNACAAGCTTGNNGCCGCACTCNAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGANAGGAANCTGAGTTGGCTGCTGCCNC
CGCTGANCAATAACTANCATAACCCCTTGNGGCCTCTANACGGGNCTTGAGGGGTTTTTTGCTGAAANGANGAACTATATCNNGATTGGCNAATGGGACGCGCCCTGTANCNGCG
CATTAAGCGCGGCGGGTGTGGTGGTTACGCGCANCGTGACCGCTACACTTGCCAGCGCCCTANCGCCCGCTCCTTTCNCTTTCNTNNCCTTCCTTNCTCGNCCACGTTCGCCNGGCT
TTNGNNNNTNNAGCTCNNAATCNGNGNGGCTCCNNNTANNNNNCCGATTNNNGCTTTACNNCACNNCNACCNNNNNNNACTTGANTAGNGNNGATGGTNCNNGTAGNNNG
NNCATCNNCCCNGANNNNNCGGNNNNNN

[GGSHVP]24:
NNNNNNNNNNNNNNNTCCNNCTNNNNNNATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGGCTCTCATGTGCCTGGAGGATCGCATGTCCCTGG
GGGTTCACACGTTCCGGGCGGTAGTCATGTACCAGGTGGCTCTCATGTGCCTGGAGGATCGCATGTCCCTGGGGGTTCACACGTTCCGGGCGGTAGTCATGTACCAGGTGGCTCTCAT
GTGCCTGGAGGATCGCATGTCCCTGGGGGTTCACACGTTCCGGGCGGTAGTCATGTACCAGGTGGCTCTCATGTGCCTGGAGGATCGCATGTCCCTGGGGGTTCACACGTTCCGGGC
GGTAGTCATGTACCAGGTGGCTCTCATGTGCCTGGAGGATCGCATGTCCCTGGGGGTTCACACGTTCCGGGCGGTAGTCATGTACCAGGTGGCTCTCATGTGCCTGGAGGATCGCAT
GTCCCTGGGGGTTCACACGTTCCGGGCGGTAGTCATGTACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCGAG
CTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAA
CTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCG
GGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCNGCTTTCCCCGTCAAGCTCTAAA
TCGGGGGCTCCCTTTAGGGTTCCGATTTANTGCTTTACGGCACNCGACCCCNAAAAACTTGANTAGGGTNATGGNTCNCGTAGTGGGGCNTTCGCCCTGNANANANNNNTTTTTCN
NCNTTGNNGNTGGNANTTCANGTTCTTNNTNNNGACTCNTGNNCNANNNNNNNNNCNNTCANNCCTATCNNNGTCTATNCTTTNNANTTNNTANGNNNNTTTNNNNNNTNN
NNGNNNNNNN

[GGLSIP]28:
NNNNNNNNNNNNNNNNNNNCTAGAANTAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGGCTTGTCAATCCCTGGAGGACTTTCTATTCCGGGC
GGTCTTAGCATTCCTGGGGGTTTATCGATACCAGGTGGCTTGTCAATCCCTGGAGGACTTTCTATTCCGGGCGGTCTTAGCATTCCTGGGGGTTTATCGATACCAGGTGGCTTGTCAAT
CCCTGGAGGACTTTCTATTCCGGGCGGTCTTAGCATTCCTGGGGGTTTATCGATACCAGGTGGCTTGTCAATCCCTGGAGGACTTTCTATTCCGGGCGGTCTTAGCATTCCTGGGGGTT
TATCGATACCAGGTGGCTTGTCAATCCCTGGAGGACTTTCTATTCCGGGCGGTCTTAGCATTCCTGGGGGTTTATCGATACCAGGTGGCTTGTCAATCCCTGGAGGACTTTCTATTCCG
GGCGGTCTTAGCATTCCTGGGGGTTTATCGATACCAGGTGGCTTGTCAATCCCTGGAGGACTTTCTATTCCGGGCGGTCTTAGCATTCCTGGGGGTTTATCGATACCAGGTGAAAACC
TGTATTTTCAGGGCCATCATCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGA
GATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGANTTGGCTGCTGCCACCGCTGANCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAA
GGAGGAACTATATCCNGANTGGCGAATGGGNCGCGCCCTGTANCGGCGCNNNAAGCGCGNCGGGTGTGGTGGTTACGCNNANCGTGACCNNNTACACTNGCCAGCGCNCTAGC
GCCCGCTCCTTTCGCTTTCTTCCNNNNNTTTCTCGNCACGTTCGCNNNTATNCCCNTCAAGCTCTAATCGGGGGNNNCCTTTAGGGNNNNANTTNNTGNTTTACGGCNCNNNNCC
CNAAAACTTGANNNNNNATGNNACGTANNNGNCCATCNNCNNTGNNNNNNNNN

[GKFGTP]24:
NNNNNNNNNNNNNNNNNNNCTNNNNNNATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTAAATTCGGTACTCCAGGAAAGTTTGGGACGCCCGGA
AAATTTGGAACCCCTGGCAAGTTCGGCACACCAGGTAAATTCGGTACTCCAGGAAAGTTTGGGACGCCCGGAAAATTTGGACCCATGGCAAGTTCGGCACACCAGGTAAATTCGG
TACTCCAGGAAAGTTTGGGACGCCCGGAAAATTTGGAACCCCTGGCAAGTTCGGCACACCAGGTAAATTCGGTACTCCAGGAAAGTTTGGGACGCCCGGAAAATTTGGAACCCCTG
GCAAGTTCGGCACACCAGGTAAATTCGGTACTCCAGGAAAGTTTGGGACGCCCGGAAAATTTGGAACCCCTGGCAAGTTCGGCACACCAGGTAAATTCGGTACTCCAGGAAAGTTT
GGGACGCCCGGAAAATTTGGAACCCCTGGCAAGTTCGGCACACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTC

GAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAA
TAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCG
GCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCT
AAATCNGGGGCTCCCTTTAN

[GSIPVPGSIPVPGSMPVPGSIPV]6:
NNNNNNNNNNNNNNNNNNCTAGAATAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCTGGTTCCATACCAGTCCCAGGTTCCATTCCAGTACCGGGTTCC
ATGCCTGTTCCTGGCTCCATCCCTGTACCTGGTTCCATACCAGTCCCAGGTTCCATTCCAGTACCGGGTTCCATGCCTGTTCCTGGCTCCATCCCTGTACCTGGTTCCATACCAGTCCCA
GGTTCCATTCCAGTACCGGGTTCCATGCCTGTTCCTGGCTCCATCCCTGTACCTGGTTCCATACCAGTCCCAGGTTCCATTCCAGTACCGGGTTCCATGCCTGTTCCTGGCTCCATCCCT
GTACCTGGTTCCATACCAGTCCCAGGTTCCATTCCAGTACCGGGTTCCATGCCTGTTCCTGGCTCCATCCCTGTACCTGGTTCCATACCAGTCCCAGGTTCCATTCCAGTACCGGGTTC
CATGCCTGTTCCTGGTTCCATCCCTGTACCTGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCGAGCTCCGTCGACAA
GCTTGCGGCCGCACTCGAGCACCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCGGAGCAATAACTAGCATAACCC
CTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGNNNGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTA
CGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTC

[GMPITPGIPITPGIPMTPGMPIT]6:
NNNNNNNNNNNNNNNNNNNNNCTAGANTAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGTAATTGGTATGCCTATAACCCCAGGGATTCCTAT
CACTCCTGGAATACCGATGACGCCCGGAATGCCAATTACACCTGGTATGCCTATAACCCCAGGGATTCCTATCACTCCTGGAATACCGATGACGCCCGGAATGCCAATTACACCTGG
TATGCCTATAACCCCAGGGATTCCTATCACTCCTGGAATATCGATGACGCCCGGAATGCCAATTACACCTGGTATGCCTATAACCCCAGGGATTCCTATCACTCCTGGAATACCGAT
GACGCCCGGAATGCCAATTACACCTGGTATGCCTATAACCCCAGGGATTCCTATCACTCCTGGAATACCGATGACGCCCGGAATGCCAATTACACCTGGTATGCCTATAACCCCAGG
GATTCCTATCACTCCTGGAATACCGATGACGCCCGGAATGCCAATTACACCTGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATC
CGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCT
GANCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTA
AGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGNCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCA
AGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTANTGCTTTANGGCACNCNNNCCCNAAAACTTGATTNNGTGATGGNTCACGTANGGGNCATCGCCTGANANACGNTTTCG
CCNNGACNNNNNTCNNNNTNCTTTAANANTNGNNNNTGNTCCAAACTGGNNANNN

[GGSQIP]24:
NNNNNNNNNANNNTNNNNNCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGGCAGTCAGATACCGCGCGGATCACAAATTCCATG
GGGTTCTCAAATCCCAGGAGGCTCGCAAATACCAGGTGGCAGTCAGATACCGCGCGGATCACAAATTCCAGGGGGTTCTCAAATCCCAGGAGGCTCGCAAATACCAGGTGGCAGT
CAGATACCGCGCGGATCACAAATTCCAGGGGGTTCTCAAATCCCAGGAGGCTCGCAAATACCAGGTGGCAGTCAGATACCGCGCGGATCACAAATTCCAGGGGGTTCTCAAATCC
CAGGAGGCTCGCAAATACCAGGTGGCAGTCAGATACCGCGCGGATCACAAATTCCAGGGGGTTCTCAAATCCCAGGAGGCTCGCAAATACCAGGTGGCAGTCAGATACCGCGCGG
ATCACAAATTCCAGGGGGTTCTCAAATCCCAGGAGGCTCGCAAATACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGT
ATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGA
NCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAG
CGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCNGCTTTCCCCGTCAAG
CTCTAAATCNGGGGCTCCCTTTNGGGTTCCNATTTAGTGCTTTANNNACCTCNACCCCAAAAACTNATTAGGTGANGGTTCACGTANNGGGCCATCNCCCTGANANACGGTTTTNG
CCCTTTGACNTGNANTCCANGTTCTNTANNNNNGNNNNCTNG

[GIRGNPRLRGTPEYPGTPGIRGTP]7:
NNNNNNNNNNNNNNNTNNNNCTANNNTNATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTATAAGAGGTAACCCCAGGCTTAGAGGGACGCCCGA
GTATCCGGGTACTCCTGGAATACGAGGTACACCAGGTATAAGAGGTAACCCCAGGCTTAGAGGGACGCCCGAGTATCCGGGTACTCCTGGAATACGAGGTACACCAGGTATAAGA
GGTAACCCCAGGCTTAGAGGGACGCCCGAGTATCCGGGTACTCCTGGAATACGAGGTACACCAGGTATAAGAGGTAACCCCAGGCTTAGAGGGACGCCCGAGTATCCGGGTACTC
CTGGAATACGAGGTACACCAGGTATAAGAGGTAACCCCAGGCTTAGAGGGACGCCCGAGTATCCGGGTACTCCTGGAATACGAGGTACACCAGGTATAAGAGGTAACCCCAGGCT
TAGAGGGACGCCCGAGTATCCGGGTACTCCTGGAATACGAGGTACACCAGGTATAAGAGGTAACCCCAGGCTTAGAGGGACGCCCGAGTATCCGGGTACTCCTGGAATACGAGGT
ACACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCA
CCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGA
GGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTG
CCAGCGCCCTAGCGCCCGCTCCTTTCGCTTNCTTCCNNTTCCTTTCTCGCCACGTTCGCCGNN

[GGLQQP]24:
NNNNNTNNNNNNNTTNNCNNGNGNGTTTNNNNNNNAAACNNGGGCTNGNCTNGNTNNNNNNNNNGGNAANNNTNNNNNAGNANNNNNNNGCNNNNTNNGNGACNTNG
TNNNACGNANNGNTTTCACNTTNNNNNNGANNTGACTNTNTNCGGNCGCTNTCATGCCATNNCNNGAAANGTTTGCGNCATTCGATGNNNTNNGGGATCTNGACGNTNTCCTT
ATGCGACTCNTGCATTAGGAAGCAGCCCAGTAGTAGGTTGAGGCCGTTGAGCACCGCCGCCGCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCACGGGG
CCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGT
GATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTT
GTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGGACTCCAACAGCCAGGAGGTTTACAACAACCCGGCGGTCTTCAGCAGCCTGGGGGATTACAGCAACCAG
GTGGACTCCAACAGCCAGGAGGTTTACAACAACCCGGCGGTCTTCAGCAGCCTGGGGGATTACAGCAACCAGGTGGACTCCAACAGCCAGGAGGTTTACAACAACCCGGCGGTCT
TCAGCAGCCTGGGGGATTACAGCAACCAGGTGGACTCCAACAGCCAGGAGGTTTACAACAACCCGGCGGTCTTCAGCAGCCTGGGGGATTACAGCAACCAGGTGGACTTCAACAG
CCAGGAGGTTTACAACAACCCGGCGGTCTTCAGCAGCCTGGGGGATTACAGCAACCAGGTGGACTTCAACAGCCAGGAGGTTTACAACAACCCGGCGGTCTTCAGCAGCCTGGGG
GATTACAGCAACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTC
GAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGNNNNNNNNNNNNNN

[GMGVAP]24:
NNNNNNNNNNNNNNNNNNNNCCCNTCTTNNNNNNATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTATGGGCGTAGCTCCAGGGATGGGGGGTT
GCTCCTGGCATGGGTGTCGCTCCGGGAATGGGAGTAGCACCAGGTATGGGCGTAGCTCCAGGGATGGGGGGTTGCTCCTGGCATGGGTGTCGCTCCGGGAATGGGAGTAGCACCA
GGTATGGGCGTAGCTCCAGGGATGGGGGTTGCTCCTGGCATGGGTGTCGCTCCGGGAATGGGAGTAGCACCAGGTATGGGCGTAGCTCCAGGGATGGGGGTTGCTCCTGGCATGGG
TGTCGCTCCGGGAATGGGAGTAGCACCAGGTATGGGCGTAGCTCCAGGGATGGGGGTTGCTCCTGGCATGGGTGTCGCTCCGGGAATGGGAGTAGCACCAGGTATGGGCGTAGCTC
CAGGGATGGGGGTTGCTCCTGGCATGGGTGTCGCTCCGGGAATGGGAGTAGCACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAG
GATCCGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGANANCCGGCTGCTAACAAAGCCCGAAAAGGAAGCTGAGTTTGGCTGCTGC
CACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTTGAGGGGGTTTTTTGCTGAAAGGAAGGAACTATATCCNGATTTGGCGAATGGGGACGCGCCCTGN
ANNNNCGCATTAANCGCGGNCGGGTGTGGTGTTTNCACNCANCGTGACCGCTACACTTTGCCANCNCCCTAACGCCCCGCTCCTTTNNN

[GNPNSPGRPSSPGSPSSPGKPSS]6:
NTNNNNNNTNTTNNNCNNNNNTTTGNNNGNANNNGNNNGCNNNNNNNNNNNNNNGNANNNNNNGNNANNNNNNNNCNGCNNNNTNNGCGACATNGNNNNNNTTANNN
NTNNNNNTCNNNANCNGNANNGANNNNCTTNCNGGGCGCTNNCATGCCATNNCGNGANNGNTTTGNNNCATTCGATGNNNTCCGGNNTCTNGACGCTCTCCNTTATGCGACT
CCTGCATTAGGAAGCAGCCCAGTAGTAGGTNGAGGCNGTTGAGCACNGCCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACANTCCCCCGGCCACGGGNCCTGCCA
CCATACCCACGCCGAAACAAGCGNTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCC
GGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTTAAC
TTTAAGAAGGAGGAGTACATATGAGCAAAGGACCTGGTAACCCTAATTCCCCAGGAAGACCGAGTTCTCCGGGTAGTCCAAGTTCGCCTGGCAAACCTAGCTCACCTGGTAACCCT
AATTCCCCAGGAAGACCGAGTTCTCCGGGTAGTCCAAGTTCGCCTGGCAAACCTAGCTCACCTGGTAACCCTAATTCCCCAGGAAGACCGAGTTCTCCGGGTAGTCCAAGTTCGCCT
GGCAAACCTAGCTCACCTGGTAACCCTAATTCCCCAGGAAGACCGAGTTCTCCGGGTAGTCCAAGTTCGCCTGGCAAACCTAGCTCACCTGGTAACCCTAATTCCCCAGGAAGACC
GAGTTCTCCGGGTAGTCCAAGTTCGCCTGGCAAACCTAGCTCACCTGGTAACCCTAATTCCCCAGGAAGACCGAGTTCTCCGGGTAGTCCAAGTTCGCCTGGCAAACCTAGCTCACC
TGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACC
ACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAANGNNNNNNNNNNNNN

[GDGYAP]24:
NGNNNNTNNTNTNNNNGNCNTCGTATAACNTNNNNNNTTTCACATTNACCNNNCNNNANTNACTNTNTNCGGNNGNNATCATGCCATACCNNGAANGTTTTGNGCCATTNG
ATGNTGTCCGGGATCTCGACGCTNTNNNTATGCGACTCCTGCATTAGGAAGCANNNAGTAGTAGGTTGAGGCNGTTGAGCACNGCCGCCGCAAGGAATGGTGCATGCAAGGAGA
TGGCGCCCAACAGTCCCCNGGCCACGGGNCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCGATATA
GGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGC
GGATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGATGGTTATGCCCCCTGGCGATGGATATGCTCCTGGAGACG
GCTACGCTCCCGGAGATGGGTATGCACCAGGTGATGGTTATGCCCCTGGCGATGGATATGCTCCTGGAGACGGCTACGCTCCCGGAGATGGGTATGCACCAGGTGATGGTTATGCCC
CTGGCGATGGATATGCTCCTGGAGACGGCTACGCTCCCGGAGATGGGTATGCACCAGGTGATGGTTATGCCCCTGGCGATGGATATGCTCCTGGAGACGGCTACGCTCCCGGAGAT
GGGTATGCACCAGGTGATGGTTATGCCCCTGGCGATGGATATGCTCCTGGAGACGGATACGCTCCCGGAGATGGGTATGCACCAGGTGATGGTTATGCCCCTGGCGATGGATATGCT

CCTGGAGACGGNTACGCTCCCGGAGATGGGTATGCACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCGAGCTC CGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAANGNANNNNNNNNNNN
[GGFMQP]28: NNNNNNNNNNNNNNNCTAGAATAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGGATTTATGCATCCTGGCGGTTTTATGCAACCCGGAGGATTC ATGCAACCTGGGGGCTTTATGCAACCAGGTGGATTTATGCANCCTGGCGGTTTTATGCAACCCGGAGGATTCATGCAACCTGGGGGCTTTATGCAACCAGGTGGATTTATGCANCCT GGCGGTTTTATGCAACCCGAAGGATTCATGCAACCTGGGGGGCTTTATGCAACCAGGTGGATTTATGCATCCTGGCGGTTTTATGCAACCCGGAGGATTCTTGCAACCGGGGGGCTT TATGCAACCAGGTGGATTTATGCAGCCTGGCGGTTTTATGCAACCCGGAGGATTCATGCAACCTGGGGGCTTTATGCAACCGGGTGGATTTATGCATCCTGGCGGTTTTATGCAACCC GGAGGATTCATGCAACCTGGGGGCTTTATGCAACCAGGTGGATTTATGCAGACTGGCGGTTTTATGCAACCCGGAGGATTCATGCAACCTGGGGGCTTTATGCAACCAGGTGAAAA CCTGTATTTTCNNGGCCATCACCATCACCATCACGGCTAATGATGATCTTCNNGATCCNNATTCNANCTCCGTCNACGANCTTGCGGCCGCACTCNAGCACCACCACCACCACCAC TGANATCCGGCTGCTAACANANCCNNANAGGAANCTGANTTGGCTGCTGCCNCCGCTGANCAATAACTANCATAACCCCTTGGGGCCTCTAAACGGNTCTTGATGGGTTTTTTGCT GAAANGANGAACTATATCCGGATTGGCNAATGGNACGCGCCCTGTACCGGCGCATTAAGCNCGGCGNNTGTGNTGNTTNNGCACANCGTGACCGCTACACTTNNCNGCGCCCTN ACGCCCGCTCCTTTCACTTTCNTCNCNTTCTTTCTCNCCACGNTNCCNNNCTGTCCCCGNNAANNNCNNNATCGGGGNCNCCNNN
[GDNGSP]24: NNNNNNNNNNNNNNNNNNNNNCTNNNNNNATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGACAATGGCAGTCCCGGCGACAACGGTAGCCCTG GGGATAACGGAAGTCCTGGAGATAATGGTTCACCAGGTGACAATGGCAGTCCCGGCGACAACGGTAGCCCTGGGGATAACGGAAGTCCTGGAGATAATGGTTCACCAGGTGACAA TGGCAGTCCCGGCGACAACGGTAGCCCTGGGGATAACGGAAGTCCTGGAGATAATGGTTCACCAGGTGACAATGGCAGTCCCGGCGACAACGGTAGCCCTGGGGATAACGGAAGT CCTGGAGATAATGGTTCACCAGGTGACAATGGCAGTCCCGGCGACAACGGTAGCCCTGGGGGATAACGGAAGTCCTGGAGATAATGGTTCACCAGGTGACAATGGCAGTCCCGGC GACAACGGTAGCCCTGGGGATAACGGAAGTCCTGGAGATAATGGTTCACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATC CGTATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCT GAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTA AGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTANCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCA GCTCTAAANCGGGGGCTCCCTTTANGGTTCCGANTTANTGCTTTANNGCACTCGACCCNNAAAACTTGATTANNTGATGGTTCACGTAGTGGCCATCGCCCTGATAGACGGTTTTTC NCCCTTTGACGTTGGAN
[GVGYAP]24: NNNNNNNNNNNNNNNNNNNCTAGAATAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTGTAGGCAATGCGCCAGGAGTAGGATATGCCCCTGGA GTCGGTTACTCTCCCGGAGTTGGGTATGCACCAGGTGTAGGCAATGCGCCAGGAGTAGGATATGCCCCTGGAGTCGGTTACTCTCCCGGAGTTGGGTATGCACCAGGTGTAGGCAAT GCGCCAGGAGTAGGATATGCCCCTGGAGTCGGTTACTCTCCCGGAGTTGGGTATGCACCAGGTGTAGGCAATGCGCCAGGAGTAGGATATGCCCCCTGGAGTCGGTTACTCTCCCG GAGTTGGGTATGCACCAGGTGTAGGCAATGCGCCAGGAGTAGGATATGCCCCTGGAGTCGGTTACTCTCCCGGAGTTGGGTATGCAACAGGTGTAGGCAATGCGCCAGGAGTAGGA TATGCCCCTGGAGTCGGTTACTCTCCCGGAGTTGGGTATGCACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCG AGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAAT AACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGG CGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTA AATCGGGGGCTCCCTTTAGGGTTCCGANTTAGTGCTTTACGGCACNCGACCCCAAAAACTTGATTANGGTGATGGTTCACGTANTGGGCCATCGCCCNGATAGACGGTTTTTCNNCC CNTTGACGTTGNAN
[GNIGTP]28: NNNNNNNNNNNNNNNNNNNNNNNNNNTNNNNNTAGAATAATTTTGTTTAACTTTAAGAAGGAGGAGTACATATGAGCAAAGGACCAGGTAATATCGGTACGCCCGGAAACAT AGGAACCCCTGGCAACATTGGGACTCCTGGAAATATAGGCACACCAGGTAATATCGGTACGCCCGGAAACATAGGAACCCCTGGCAACATTGGGACTCCTGGAAATATAGGCACA CCAGGTAATATCGGTACGCCCGGAAACATAGGAACCCCTGGCAACATTGGGACTCCTGGAAATATAGGCACACCAGGTAATATCGGTACGCCCGGAAACATAGGAACCCCTGGCA ACATTGGGACTCCTGGAAATATAGGCACACACCAGGTAATATCGGTACGCCCGGAAACATAGGAACCCCTGGCAACATTGGGACTCCTGGAAATATAGGCACACCAGGTAATATC GGTACGCCCGGAAACATAGGAACCCCTGGCAACATTGGGACTCCTGGAAATATAGGCACACCAGGTAATATCGGTACGCCCGGAAACATAGGAACCCCTGGCAACATTGGGACTC CTGGAAATATAGGCACACCAGGTGAAAACCTGTATTTTCAGGGCCATCACCATCACCATCACGGCTAATGATGATCTTCAGGATCCGTATTCGAGCTCCGTCGACNAGCTTGCGGCC GCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGANCAATAACTAANCATAACCCCTTGGGGCC TCTAAACNGTCNTTGAGGGGTTTTTTGCTGAANNGANGAACTATATCCCGGATTGGCGATGGGACGCGCCCTGNAGCNGGCGCATTAAGCGNGNCNNGTGTGGNGGNTACGCGCA GCGTGACCGCTACACTTGCCAGCGNCCTANCGNCCGGCTCCTTTNGNTTTCTTCCNTNCNTTNTCGNCACGTCGCCGGNNTTNNNCGTCAAGCTCNAATCGGGGGCTCCTTNNNGN TCNNANTNANNNTTTACGNNACCTCGNNCCNNAAANNTTNNTNNNGTGNNTGNTTNN

Table 14: Nonrepetitive sequences that display LCST behavior

A[0.2] rep, 200 amino acids: $(GVGVPGVGVPGAGVPGVGVPGVGVP)_{8}$
A[0.5] rep, 200 amino acids: $(GAGVPGVGVP)_{20}$
A[0.2] rep, 400 amino acids: $(GVGVPGVGVPGAGVPGVGVPGVGVP)_{16}$
A[0.5] rep, 400 amino acids: $(GAGVPGVGVP)_{40}$
A[0.2] nonrep, 200 amino acids: PGGGVPGGVPGAVPGVPGAVVPVVGGVPGVVGVPGGVVPVGVVVPGVPVGVPGVPVVGGGVPAGGVVPGGGVPVGGGVVPGVGVVPGGVPVVVGGGVGVPGVVPGVPGVGVPVGGVP GGGVVVPGGVGVPVVGVVPGAPGVPGVPGGVGGVPVGVGVPGGGAPGVVPGGAVPGGAPVGGVGGGVPGVGGVPGAPGGVPG
A[0.5] nonrep, 200 amino acids: PGGVPGVPAGGGAGVPGVPGAPAGGGAPGVAPGGVVPGVPGVAPVVGGGAPGGVPGAVPGVPGGVPGGVVVPGGGVGAPGGVAPVGGGVPVVGGVPVGGVAPGVPVGVPGVGVGVG VPVGGGVGGVPGVVPAGVPGGGVAGVVPVGGVPVVVGAPGGGVGVPGAPGAGGVPGGAPGAPGVVAPGGVPVGVVVPGVVPGGG
A[0.2] nonrep, 400 amino acids: PGGVPGVGGGVPGVPGAPGVVPGVVVPGGVVPVGGVPGVPVVGGGVPVGGGVGGVPGAPGGGVPGAPVGGVPVGGVVPGVGVPGAGGGVPGGVGVVPVGGGAPGGVPGVVPGVPVGG VPAGVVPGGVGVGGGVVPGVVGVPGGGVGGGVGVPGGAPGGGVVGGGVPGGGVVPVVGGVPGAVPGVPGVVVPVGGGVVPVGVPGGVVVPGVPGGVVGVPVVGVPGVGVPVGPVGPVP VGVGVVVPGVVPVGVGVPGVPGGGVPGVGVPVGGAPGVPGVPGVAVPGVVPGGAPGVPVGVVPVGVVPGGGVGVVPGGVPGGVPGVPAGVPVGVPVGGVGGVVPGGGAPGVGVGVPA GVPGGVPVGVGGVPGGVVPGVPAGGGVPVGVPGVGGVPVGAPGVPGGVP
A[0.5] nonrep, 400 amino acids: PAGAVPGGVPVAGGVPVGAGVVAPGAPVGGAPGVVGVPGGGVPGVPVVGVGVGVPVGGGVGVVPAGGGAPGGVGVVVPGVVPGAPAGGVPGVAPGGVGGVPGVVPGGGVVPGVAPG VPGVPGGVPVGVPVGAAPGGGVPAVGVPGVGGGVPVVGAGAPGGVPGVAVPGAPGGVPVGGVPGGGVGGVVPGGVPGAPVGVVPGVGVVPGVPAGVGVPGAPGAPGGGVPGGGAPVG GAGGVPGGVPGVPGAPGVGVPGVGVPVGVPGVPGVPVGGGAGVPGGVAPGVAVPGGVGVPVVGGVPGAVPGVPGVGGAPGVPGGGVVGVGVPGGVVPGVPGAGAPGGGAPGAPVGVP GGAPGVVPGVGGVPGAGGGVVGGGVVPVGGGVVPGAPGGAPVGGVVGGVP
VVA: PVGVPGVGGAPVVVGGAPVGVGGAGGVVPGAPVGAVAGVGAPGGAVPAGGAPGVGAPGVVPGAVPGVGVGGVGGGVVVPVVVGVGGVPAGAVPVGGVAPVGVPVGAGGGAPAGGV PVAVGVPVVGGGVPGVGVVGVAPGAPGAGVGVVAGAGAAAVAPVGVAPGVPGVPGAVVGVPGGVAPAVVVGAVPGVVGVVPVGAPVVAGGVAVAPVGGAVPGVPGGVPGGGAPGAP VVGVPVGVP
VVS: PVGSPGSVPGVVPVSGVVVGGGSVPGSSPVVGVPSGVPVGSPVVGGGVSPGVGVPGVVVPGVPVGGGVGVPGGGVVGGVPVVVGVSVPSGVSPGGGVVVPVGSVSSGVPVVSVVGVSVVPGS VPVGSPSSGVPGVVPGSPGGVVGSSVPSGSGGVSVPVGVGGSGVPGVPGSGSVVPVGVSPVGGGSGSGVGSGSPGGSVPGGSPGVVPVGSPGVGGVPGVVGGSGSPGVPGGVGVGVPG
VVT: PVGTPGTVPGGVVPVGVTPVGGGTGTPTTVTVVGGVVVPGVPGGGTPGVPVGTPVVGGVGVGTGTPGGTVPGGTPGVVPGTPGGVVPGTTVPTGTGTTPVVGVPTVGGVTPGVGGVPVVVG TGVPGVPVGGGVGVVGVVVPGTVPVGVPGVVGGTVPVGTPTTGVPGVVPTGVTPGGVGVPGGGVVGGVVPGVTVPVGTVTTGVPVVTVGVVVPVGTVGGVVVGGGTTPGGVPGTPGVGG
VAL: GAVVVPGGAVGVPAGAVGGVGGVLVPAVGAVPGGGVPAVGVPVAGVVPGGLPGGAGVGALGAAAPGGVPVGAALPGVAGVAPGGGLPGLGAGAGLGLAGALVLPGLGGVPGVPGGG LLPGGVPLLGLPGVPAGLPGVLPVGLLLPGAPGVPALGVPLGVPGVAPGLPGAGGLPGVGAPGLLPVGGAPAALGGLGLPGGGALGLLAPGGALVPGVGGVAPVAAGALLPGGAPLGVPAL LAL
VFS: SSGSSGSGFFSGGVPSGGGVVFVPSFFGFFPSGSGGVVPGVPGGGVGVPGFVGVPVGGFPVGVPGVFVPSVVGVPGFFVFPGVPGGSSSPGSPFGVFGGGSSGGGFPFGSSPGVVVSPGFGFGVPFG VPVGGSVFFFPFGGGFPGGFPSGSFGVPGGGFVVPFGVSPGSPGFVPSFGSPGGFGGFSPFSSGVGSPGGVPGVGVVSPGFPGGSPFGFPGSPGGSFGGSPSVFGSGSPGSFP
AFT: AAGGTGFFTGGAPGGTPTAFGGFTPFTFAGGGAPGFFAFPGAPGTFPTFGAPGAAPFGTTPGTFFGFPGGTPGTPGAATPGFFGAFGFPGAFATFGFPGGGAGAAPAGTPGAPTAAGATTPGA GGTTGGAPTGGGFAPTFFGFFPTGTGGAAPGAPGGGTTTFPGATPFGFPGTGTPGTFGAPGGGFFPGAGFGGGAPGAPAAAGGGTPGTATPGTFPTGAGTFPGTFPFGGGFPGGFPT
ILT: TILPGIGLPGGLPLTIGGILPILLGTTIGGIPGGGIIPGIGGGIPILTLLILPTGLPGTPLTTGTPGITGLGTPLGIGLIGTGIPGGGTGIPTGIPGTTLPGIPTLGGGIPLIGTPGGIGTIGIPGTGTIIGIPGGGL LLIPGITPLIGLPGGTPGTPGGGTGTTPTGLPGILLPGGLPGIPTGTPGTGTPIGGLLILLTLPGLPGLGLPGLIPLGTGLTGGTPTGGIPGGIGIPGILTT
VALKTIY: PGTYPGYGYVYPTTGGIPGGVVPGGGTKLVPGKGKGGAKAPGTVPVGAGGGKIVPIYGIAPGKYGYPGGGIVPGITTPGLPTGKKPYGGVPVLYGKLPGAPGIPTAGAPGYIAPGVPGGLVKGG TGIAPLGVILVYIGVGGIKGGALPTGGLYPGAGTPGYPVGGGAPAGGIALKPGITPGTAAPGLPGKGGKYTYPGGAPGGTGGVPGYPGLALKLGIPTKGGGIGLPYIGLLPGKPGG

Table 15: Nucleotide sequences of genes generated using codon scrambling for high throughput DNA synthesis and protein production in *E. coli*

[GGHWLP]28: GGTGGTCATTGGCTTCCTGGGGGACATTGGTTACCCGGCGGTCACTGGTTGCCTGGAGGGCATTGGTTACCAGGAGGGCACTGGTTACCGGGAGGACATTGGCTACCAGGGGGCCA CTGGTTGCCCGGAGGCCATTGGCTGCCTGGTGGCCACTGGTTACCTGGAGGTCATTGGCTCCCAGGCGGACATTGGCTTCCGGGTGGACACTGGCTCCCGGGCGGGCATTGGTTGCC AGGCGGTCATTGGCTACCTGGTGGGCACTGGCTACCCGGGGGACACTGGTTACCCGGTGGTCACTGGCTACCTGGCGGACACTGGCTGCCAGGGGGTCACTGGTTACCAGGTGGAC ATTGGCTCCCCGGAGGTCACTGGCTTCCCGGAGGACACTGGCTACCAGGCGGCCATTGGTTACCTGGGGGCCATTGGTTGCCGGGAGGCCACTGGCTACCGGGTGGGCATTGGCTTC CAGGTGGCCATTGGCTCCCTGGGGGTCATTGGTTACCG
[GRRGSP]28: GGTCGTAGAGGATCTCCAGGAAGAAGGGGTAGTCCCGGTAGGCGAGGTAGCCCTGGCAGACGTGGGTCACCAGGAAGGCGCGGTAGTCCAGGCAGAAGAGGAAGTCCTGGTAGG AGGGGATCTCCCGGACGTCGAGGAAGCCCGGGAAGAAGAGGGAGTCCAGGACGTAGGGGAAGTCCAGGTCGTCGCGGCTCCCCAGGCCGAAGGGGGTCCCCTGGACGCCGTGGA TCTCCGGGTCGGAGAGGCAGCCCCGGAAGACGGGGATCACCCGGCAGGCGTGGCTCACCTGGAAGACGCGGATCTCCTGGTCGCAGAGGATCACCGGGTAGAAGAGGCTCTCCAG GTAGAAGGGGCAGTCCTGGGCGTAGAGGTTCTCCTGGAAGGAGAGGATCGCCAGGTAGGAGAGGTTCCCCCGGTCGAAGAGGTTCGCCTGGTAGACGAGGTTCACCTGGGAGGCG GGGTTCTCCAGGGAGAAGAGGTAGTCCTGGCCGTCGTGGTAGTCCG
[GGKQAP]28: GGTGGAAAGCAGGCTCCTGGCGGTAAGCAAGCTCCCGGAGGTAAACAAGCCCCTGGAGGAAAACAGGCACCAGGAGGAAAACAAGCTCCGGGCGGAAAACAAGCGCCTGGAGG TAAGCAAGCACCTGGAGGGAAGCAGGCGCCGGGTGGGAAGCAAGCGCCAGGTGGGAAACAAGCACCAGGCGGTAAACAGGCACCTGGTGGAAAACAAGCCCCAGGAGGGAAA CAAGCTCCTGGAGGCAAGCAGGCCCCAGGGGGCAAACAAGCTCCAGGAGGCAAACAGGCTCCGGGTGGTAAACAAGCACCGGGAGGAAAGCAAGCCCCCGGAGGGAAACAGGC GCCCGGTGGTAAGCAGGCACCCGGTGGAAAACAGGCTCCCGGCGGGAAACAAGCCCCGGGAGGTAAACAGGCCCCTGGTGGCAAGCAAGCTCCTGGTGGTAAACAGGCTCCAGG TGGCAAACAAGCACCCGGGGGAAAACAAGCACCTGGGGGTAAACAAGCTCCC
[GEGWSP]28: GGTGAAGGGTGGTCTCCTGGGGAAGGTTGGAGCCCTGGCGAAGGATGGAGTCCCGGAGAAGGCTGGTCACCAGGAGAAGGTTGGTCACCAGGCGAAGGGTGGAGTCCGGGCGAA GGCTGGTCCCCGGGAGAGGGTTGGTCACCTGGAGAGGGGTGGAGCCCCGGTGAGGGGTGGAGTCCTGGCGAGGGTTGGTCGCCAGGTGAAGGCTGGAGTCCAGGAGAGGGCTGGA GTCCTGGAGAAGGGTGGTCACCCGGGGAAGGGTGGTCGCCCGGTGAAGGTTGGAGTCCTGGTGAGGGTTGGAGTCCAGGTGAGGGCTGGTCGCCTGGAGAAGGATGGTCCCCAGG GGAAGGATGGTCACCTGGGGAGGGATGGTCACCGGGTGAGGGATGGTCCCCCGGAGAGGGATGGAGCCCAGGCGAGGGATGGAGTCCAGGGGAGGGTTGGTCCCCTGGTGAAGG ATGGTCGCCGGGGGAAGGCTGGAGCCCGGGGGAGGGGTGGTCCCCT
[GANGVP]28: GGTGCGAATGGCGTACCCGGAGCTAATGGAGTCCCTGGGGCTAACGGTGTTCCAGGAGCAAATGGTGTACCAGGAGCTAACGGAGTTCCGGGGGCAAACGGGGTCCCTGGAGCGA ATGGTGTCCCGGGTGCTAACGGGGTACCTGGGGCGAACGGTGTCCCAGGTGCGAACGGCGTACCAGGTGCCAACGGGGTTCCCGGTGCTAATGGCGTCCCTGGTGCAAACGGAGTC CCCGGCGCTAATGGGGTGCCTGGAGCAAACGGTGTGCCTGGGGCCAACGGTGTACCTGGAGCCAACGGAGTACCGGGCGCCAATGGTGTTCCTGGCGCGAATGGAGTACCAGGCG CAAATGGGGTTCCTGGTGCCAATGGCGTTCCAGGGGCAAATGGAGTTCCAGGTGCAAATGGCGTGCCCGGGGCGAATGGGGTACCAGGGGCCAATGGAGTGCCAGGGGCTAATGG TGTGCCGGGAGCCAATGGGGTCCCAGGAGCGAACGGAGTGCCT
[GVIGIP]28: GGTGTAATTGGCATTCCGGGCGTCATCGGTATTCCAGGAGTGATAGGAATCCCTGGGGTTATAGGCATACCAGGAGTAATCGGTATCCCTGGTGTAATAGGGATACCAGGGGTCATT GGAATTCCTGGAGTTATAGGAATACCTGGTGTTATCGGGATTCCGGGTGTTATAGGTATACCAGGTGTTATTGGGATACCTGGAGTCATAGGTATTCCCGGAGTTATTGGCATACCTG GCGTGATCGGAATTCCAGGGGTTATTGGTATTCCTGGTGTCATTGGTATACCCGGGGTAATAGGCATTCCTGGGGTGATTGGTATACCGGGAGTAATTGGTATCCCAGGTGTAATCGG CATTCCAGGTGTCATAGGCATCCCTGGAGTAATAGGAATTCCCGGCGTTATTGGAATCCCAGGAGTTATCGGTATACCTGGGGTAATTGGAATACCAGGCGTAATAGGTATCCCCGG TGTGATAGGTATTCCGGGGGTAATCGGAATACCC
[GTTGTP]28: GGTACTACTGGTACGCCTGGGACGACCGGAACTCCTGGCACTACAGGAACCCCTGGAACAACGGGCACACCAGGAACTACAGGGACGCCGGGAACAACTGGGACTCCAGGGACA ACAGGGACCCCAGGAACAACAGGAACACCTGGGACCACAGGAACGCCAGGTACAACTGGAACACCAGGGACTACAGGTACTCCAGGTACCACAGGTACCCCAGGTACGACTGG AACCCCGGGTACTACGGGAACTCCAGGCACAACTGGCACGCCCGGGACTACTGGAACTCCCGGTACGACAGGAACTCCGGGGACAACTGGTACCCCCGGTACCACCGGTACACCA GGTACTACCGGCACTCCAGGAACCACGGGTACACCTGGAACGACTGGTACACCGGGTACGACGGGGACTCCTGGAACTACTGGGACACCTGGTACTACAGGCACCCCTGGTACCA CTGGTACTCCTGGTACAACAGGTACACCCGGAACAACCGGTACTCCC
[GGKTIP]28: GGTGGCAAAACTATCCCAGGAGGAAAAACAATTCCAGGGGGTAAAACGATTCCGGGCGGTAAGACTATACCAGGAGGTAAAACAATACCTGGTGGGAAAACTATTCCCGGGGGA AAAACCATACCAGGTGGTAAAACCATTCCAGGTGGGAAGACTATACCTGGGGGAAAGACTATTCCGGGAGGCAAAACAATACCGGGTGGTAAAACAATCCCTGGAGGAAAAACT ATACCGGGAGGTAAGACCATTCCTGGAGGGAAAACAATACCCGGAGGAAAGACAATACCAGGCGGGAAAACGATCCCTGGTGGTAAGACGATTCCCGGCGGCAAAACGATACCT GGAGGTAAAACTATACCCGGTGGTAAAACTATTCCTGGTGGAAAAACTATTCCAGGAGGCAAGACTATCCCCGGAGGGAAGACAATTCCTGGGGGTAAGACAATCCCAGGTGGAA AGACCATACCTGGCGGAAAGACGATACCAGGGGGCAAAACCATCCCT
[GMLFPPGMLLPPGILFPPGILLPP]7: GGTATGTTGTTTCCTCCTGGAATGTTATTACCCCCTGGGATATTATTTCCGCCCGGTATATTGTTGCCACCTGGGATGTTGTTCCCTCCAGGTATGTTACTACCACCTGGAATCCTTTTTC CACCTGGCATATTACTGCCACCAGGGATGTTATTCCCACCAGGCATGTTATTGCCCCCAGGAATTCTTTTCCCGCCTGGTATTTTACTCCCACCTGGTATGCTGTTTCCACCGGGGATG CTTCTGCCTCCAGGCATTTTATTTCCTCCAGGAATACTTTTGCCTCCTGGCATGCTTTTTCCTCCGGGTATGCTCCTACCTCCTGGTATCCTCTTCCCTCCTGGGATTCTCTTGCCGCCGG GAATGCTCTTTCCTCCCGGAATGTTGTTACCTCCAGGGATCTTATTTCCACCAGGTATTCTATTACCACCAGGAATGCTATTTCCCCCCGGCATGTTGCTACCGCCAGGTATACTCTTTC CACCCGGGATACTTCTACCTCCC
[GSFGIP]28: GGTTCATTTGGCATTCCGGGTAGTTTCGGTATTCCTGGATCGTTTGGAATCCCAGGCTCTTTTGGTATACCAGGAAGTTTTGGAATTCCTGGGAGTTTTGGTATCCCCGGTTCATTCGGA ATACCGGGTTCGTTCGGAATCCCTGGATCTTTTGGAATACCCGGCTCCTTTGGTATACCGGGATCATTCGGGATTCCAGGATCTTTCGGGATACCTGGAAGTTTCGGAATTCCAGGTTC CTTTGGCATACCTGGCAGCTTTGGTATTCCAGGGTCTTTTGGGATACCAGGTTCTTTTGGCATCCCAGGTAGCTTTGGAATTCCCGGGTCATTTGGTATTCCCGGATCATTTGGAATACC TGGTTCTTTCGGCATACCAGGGAGCTTCGGTATCCCTGGGTCCTTTGGAATACCAGGCAGTTTTGGTATTCCGGGCTCGTTTGGTATACCTGGTAGTTTTGGGATTCCTGGTTCCTTCGG CATTCCTGGCTCATTCGGTATACCC
[GEPERPGLPLNPGDPLNPGLPDKP]7: GGTGAGCCAGAAAGACCGGGACTACCACTTAATCCTGGAGATCCGCTCAACCCTGGGCTGCCTGATAAACCTGGAGAACCTGAACGACCAGGACTTCCTTTGAACCCCGGTGATCC ACTGAATCCTGGCCTACCTGATAAGCCAGGCGAACCAGAACGGCCTGGTCTACCTCTCAATCCAGGTGACCCTCTTAACCCGGGGCTACCCGATAAACCGGGTGAACCAGAGCGAC CTGGTTTACCTTTAAATCCCGGAGATCCATTAAATCCGGGCCTTCCAGACAAGCCGGGAGAGCCCGAGCGTCCTGGACTCCCACTCAATCCTGGGGACCCATTGAATCCAGGGCTTC CTGACAAACCAGGGGAACCGGAACGTCCAGGTCTGCCGCTTAATCCAGGAGATCCTCTAAACCCAGGTCTTCCCGACAAACCTGGCGAGCCTGAAAGACCCGGACTGCCACTAAA TCCTGGTGATCCCTTAAATCCAGGCTTACCAGATAAACCA
[GDGYAP]28: GGTGATGGTTATGCCCCTGGCGATGGATATGCTCCTGGAGACGGCTACGCTCCCGGAGATGGGTATGCACCAGGAGACGGTTACGCCCCAGGGGATGGTTATGCACCCGGTGACGG TTATGCGCCAGGTGACGGGTATGCCCCCGGTGATGGATATGCCCCAGGAGATGGATATGCACCGGGAGACGGATATGCACCTGGTGACGGATATGCGCCTGGGGATGGGTACGCAC CGGGCGATGGTTACGCTCCGGGTGATGGCTATGCGCCCGGAGACGGGTACGCTCCTGGCGACGGATACGCACCTGGGGACGGATATGCTCCGGGGGATGGCTACGCACCCGGGGA TGGATACGCCCCTGGAGATGGATACGCTCCAGGCGATGGGTATGCTCCCGGCGATGGCTATGCCCCGGGAGATGGTTACGCACCAGGGGACGGTTATGCTCCAGGTGATGGATACG CGCCTGGTGATGGGTATGCGCCGGGTGACGGCTATGCTCCT
[GYGHAP]28: GGTTATGGACATGCTCCGGGGTATGGTCATGCGCCTGGCTACGGTCACGCTCCTGGATATGGGCATGCACCAGGTTATGGTCATGCCCCCGGTTACGGACATGCTCCCGGATATGGA CATGCACCTGGGTATGGCCATGCTCCTGGTTATGGTCACGCGCCTGGATATGGCCATGCCCCGGGCTATGGTCATGCACCGGGTTATGGGCATGCGCCAGGGTATGGTCACGCCCCT GGATATGGTCATGCTCCAGGATACGGTCATGCTCCTGGCTATGGCCATGCACCCGGCTATGGGCACGCACCCGGTTATGGCCACGCTCCAGGTTATGGACACGCCCCAGGATATGGA CACGCACCAGGCTATGGACATGCCCCTGGTTACGGCCATGCGCCGGGATATGGTCACGCACCTGGATACGGCCACGCACCGGGATACGGGCACGCTCCCGGGTATGGACATGCGCC CGGATACGGACACGCTCCGGGTTACGGGCATGCTCCT
[GGYSAP]28: GGTGGGTATAGTGCCCCTGGCGGTTACTCAGCTCCCGGAGGATATTCTGCTCCAGGAGGCTACAGTGCACCAGGGGGATATAGTGCACCGGGAGGTTACTCTGCGCCGGGTGGATA CAGCGCTCCCGGGGGTTACTCCGCGCCCGGTGGTTATTCGGCTCCTGGCGGATATTCGGCCCCCGGAGGGTATAGCGCGCCTGGCGGCTATTCAGCACCAGGCGGATACTCAGCACC TGGAGGGTACTCAGCGCCTGGGGGATACTCGGCACCTGGTGGATATTCCGCTCCGGGCGGGTATTCAGCTCCTGGGGGTTATTCTGCACCCGGTGGCTATTCCGCACCAGGAGGATA

CTCTGCTCCCGGCGGTTATTCCGCCCCTGGAGGTTATAGTGCGCCAGGTGGGTACTCTGCCCCGGGTGGTTACAGTGCTCCGGGTGGCTACTCAGCCCCAGGGGGCTACTCTGCACCT GGGGGCTATTCGGCGCCTGGAGGCTATAGTGCTCCT
[GYGYAP]28: GGTTATGGGTACGCTCCTGGATACGGATATGCCCCAGGGTATGGCTATGCTCCCGGTTACGGCTACGCACCAGGATATGGCTATGCGCCCGGGTATGGGTATGCGCCTGGTTACGGA TACGCGCCTGGGTATGGTTATGCCCCCGGATACGGCTATGCCCCGGGATACGGTTACGCCCCAGGATACGGGTATGCTCCTGGTTATGGTTATGCGCCAGGATATGGGTATGCCCCT GGTTATGGCTATGCACCTGGCTATGGTTATGCACCGGGTTATGGATACGCTCCAGGTTACGGGTACGCACCGGGGTATGGATACGCACCCGGATATGGTTACGCACCTGGATATGGA TATGCACCCGGTTATGGTTACGCTCCCGGCTATGGATATGCTCCAGGCTATGGGTATGCACCAGGTTATGGATATGCGCCGGGTTACGGTTATGCTCCGGGCTATGGCTACGCTCCGG GATATGGATACGCCCCTGGATATGGTTATGCTCCT
[GVGYAP]28: GGTGTAGGCTATGCGCCAGGAGTAGGATATGCCCCTGGAGTCGGTTACGCTCCCGGAGTTGGGTATGCACCAGGTGTCGGTTATGCTCCCGGCGTAGGTTATGCGCCGGGTGTTGGC TACGCACCGGGAGTTGGATACGCTCCTGGTGTAGGGTACGCTCCAGGCGTAGGGTATGCTCCGGGCGTTGGTTACGCACCTGGCGTGGGCTATGCTCCAGGAGTGGGGTATGCCCCG GGAGTCGGCTATGCCCCAGGGGTAGGATACGCGCCTGGTGTTGGTTATGCACCGGGGGTAGGCTACGCTCCGGGTGTGGGGTACGCACCAGGGGTTGGCTATGCACCTGGAGTGGG ATACGCACCCGGGGTAGGTTACGCCCCAGGTGTTGGATATGCTCCTGGGGTGGGATATGCACCCGGTGTGGGTTATGCCCCCGGAGTCGGGTATGCGCCCCGGTGTTGGGTACGCCCC TGGCGTCGGATATGCGCCTGGGGTCGGATACGCCCCC
[GHGNRP]28: GGTCATGGAAACAGACCCGGTCACGGAAATAGGCCAGGGCATGGGAATCGTCCAGGACATGGTAATCGACCAGGCCATGGTAATCGTCCGGGACACGGAAATCGTCCTGGGCACG GAAACAGGCCCGGTCATGGAAATCGACCGGGGCATGGCAACCGTCCAGGTCATGGTAATAGGCCGGGCCATGGCAACAGACCTGGGCATGGTAATCGGCCTGGACATGGCAATAG ACCCGGCCATGGAAACCGACCAGGACACGGTAATCGCCCTGGACACGGCAATCGTCCCGGGCATGGAAATCGCCCGGGTCATGGGAATAGGCCTGGCCACGGAAACCGGCCCGG ACACGGGAATCGACCCGGACATGGAAATAGACCTGGTCATGGCAATCGACCTGGCCATGGGAACAGACCGGGACATGGGAATAGACCAGGGCACGGTAACAGACCAGGTCACGG GAACCGTCCTGGTCACGGTAACCGCCCAGGCCACGGTAATAGACCG
[GGHNIP]28: GGTGGGCATAATATCCCTGGCGGTCACAATATACCCGGAGGTCATAACATTCCTGGAGGACACAACATACCAGGCGGACATAACATTCCCGGTGGACATAATATTCCTGGGGGACA CAATATACCGGGTGGTCACAATATTCCTGGTGGTCATAATATCCCCGGAGGACATAACATACCGGGAGGACATAATATCCCGGGCGGGCATAACATACCAGGAGGTCATAATATTC CCGGCGGCCATAATATTCCGGGAGGTCACAACATCCCAGGAGGGCACAATATACCAGGTGGGCACAACATTCCAGGGGGTCATAATATACCAGGGGGCCATAATATCCCAGGTGG ACATAACATCCCTGGAGGGCATAATATTCCAGGAGGCCACAATATACCTGGGGGTCACAATATCCCTGGTGGACACAATATTCCAGGTGGTCATAACATACCTGGTGGCCATAAACA TACCCGGGGGACATAATATACCTGGAGGCCATAATATACCC
[GLNGVP]28: GGTCTTAATGGAGTGCCTGGACTAAATGGTGTTCCGGGTTTGAATGGCGTCCCAGGATTAAACGGTGTACCAGGGTTAAACGGGGTCCCTGGACTTAATGGCGTGCCTGGCTTAAAT GGAGTCCCAGGCCTCAACGGGGTTCCCGGACTCAACGGAGTCCCTGGGCTCAACGGTGTCCCCGGTTTAAATGGCGTTCCAGGTTTAAACGGCGTACCAGGATTGAACGGTGTTCCA GGGCTTAATGGTGTACCTGGATTAAATGGGGTACCTGGTTTGAACGGAGTACCAGGTCTAAACGGAGTTCCTGGATTGAATGGGGTTCCTGGTCTGAACGGGGTACCAGGCTTGAAT GGAGTTCCAGGACTCAATGGAGTACCTGGGTTGAATGGTGTGCCAGGACTTAACGGTGTGCCGGGGCTCAATGGTGTCCCTGGCCTTAATGGGGTCCCGGGCCTTAACGGAGTGCCC GGCCTCAATGGCGTACCGGGTCTTAACGGGGTGCCT
[GYGNTP]28: GGTTATGGAAACACGCCAGGATACGGCAATACTCCGGGCTATGGTAATACCCCTGGATATGGGAATACACCAGGATATGGAAATACGCCTGGGTACGGAAATACACCTGGATATG GTAACACACCTGGTTATGGGAATACCCCCGGATACGGGAATACTCCTGGTTACGGGAACACTCCAGGATATGGTAATACGCCGGGTTACGGAAATACTCCAGGCTATGGAAATACC CCGGGATACGGAAACACCCCAGGGTATGGTAATACTCCTGGGTATGGGAATACGCCCGGGTATGGAAATACTCCCGGCTATGGGAACACACCAGGGTACGGTAATACACCTGGCT ACGGTAATACTCCAGGTTACGGTAATACCCCAGGTTATGGAAATACACCGGGTTATGGTAATACACCCGGATATGGCAACACTCCGGGATATGGAAACACTCCTGGATACGGTAAC ACTCCCGGTTATGGTAACACCCCTGGCTATGGCAATACACCT
[GGHQQP]24: GGTGGGCACCAACAGCCTGGAGGACATCAACAACCCGGCGGTCATCAGCAGCCAGGGGGACACCAGCAACCAGGAGGGCACCAGCAACCTGGAGGCCACCAGCAACCCGGAGG GCATCAACAACCAGGGGGTCATCAACAACCGGGAGGTCATCAGCAACCGGGTGGACACCAACAGCCAGGTGGTCATCAACAGCCTGGTGGACATCAACAGCCGGGCGGACATCA GCAACCTGGGGGCCATCAACAACCTGGCGGGCATCAGCAGCCTGGGGGACATCAGCAGCCGGGAGGCCATCAGCAACCAGGCGGCCACCAACAACCTGGTGGGCATCAACAGCC CGGAGGTCACCAACAACCAGGTGGCCATCAACAGCCAGGAGGACACCAACAACCCGGGGGTCACCAGCAGCCCGGTGGTCACCAACAGCCC
[GGQHIP]28v2: GGCGGCCAACACATACCTGGTGGACAACATATTCCCGGTGGTCAACATATACCCGGAGGACAACACATCCCAGGTGGACAGCATATTCCTGGAGGACAGCACATCCCTGGTGGGC AACACATACCAGGGGGTCAGCATATACCAGGTGGTCAGCATATCCCAGGGGGACAGCATATACCTGGAGGTCAACATATCCCGGGCGGTCAACATATTCCTGGTGGCCAGCACAT ACCAGGCGGTCAGCATATTCCAGGTGGCCAACATATACCAGGAGGACAACATATACCTGGGGGTCAACACATTCCAGGAGGTCAGCACATTCCTGGCGGACAACACATTCCCGGC GGGCAACATATCCCAGGAGGCCAACATATCCCCGGAGGTCAACACATACCGGGTGGGCAGCACATACCCGGGGGACAACATATCCCTGGAGGGCAACATATTCCAGGGGGCCAA CATATTCCGGGAGGGCAGCATATACCGGGAGGCCAGCATATACCC
[GVMGQP]24: GGTGTTATGGGTCAGCCAGGAGTCATGGGACAACCCGGCGTAATGGGGCAGCCTGGAGTGATGGGCCAACCAGGAGTGATGGGTCAGCCGGGTGTAATGGGCCAGCCAGGTGTGA TGGGTCAACCCGGAGTAATGGGTCAGCCTGGTGTAATGGGACAGCCTGGCGTCATGGGCCAACCCGGTGTTATGGGACAACCGGGCGTTATGGGCCAACCGGGAGTGATGGGACA GCCCGGAGTCATGGGGCAACCGGGTGTCATGGGACAGCCAGGGGTAATGGGACAACCTGGTGTTATGGGGCAACCTGGAGTTATGGTCAACCAGGGGTTATGGGACAGCCGGGA GTCATGGGTCAACCGGGGGTGATGGGGCAACCAGGCGTGATGGGACAACCAGGTGTAATGGGTCAACCTGGGGTCATGGGTCAGCCC
[GVEGIP]28: GGTGTTGAAGGCATTCCGGGCGTAGAAGGAATCCCAGGGGTCGAAGGTATTCCAGGAGTAGAGGGTATACCAGGTGTCGAGGGAATCCCTGGTGTCGAAGGGATACCTGGGGTAG AGGGGATTCCGGGAGTAGAAGGAATTCCGGGGGTTGAAGGTATACCGGGAGTCGAGGGCATCCCAGGAGTTGAAGGAATTCCAGGGGTGGAAGGCATCCCCGGTGTAGAAGGAA TACCTGGCGTCGAAGGAATACCGGGTGTGGAGGGTATTCCTGGAGTAGAAGGTATACCTGGAGTGGAAGGTATACCCGGAGTGGAGGGCATTCCCGGCGTGGAAGGGATTCCAGG TGTTGAGGGAATTCCTGGTGTAGAGGGAATACCAGGCGTTGAGGGTATCCCTGGAGTTGAGGGGATACCAGGAGTCGAAGGCATACCAGGGGTAGAAGGTATCCCAGGTGTAGAA GGTATTCCCGGGGTTGAGGGCATACCTGGTGTGGAAGGAATACCA
[GEGKVP]28: GGTGAGGGTAAAGTTCCAGGAGAAGGAAAAGTACCGGGCGAAGGCAAAGTCCCAGGGGAAGGTAAGGTACCAGGTGAGGGCAAAGTGCCTGGAGAAGGGAAAGTCCCTGGAGA AGGTAAAGTGCCGGGAGAAGGTAAGGTGCCAGGGGAGGGAAAAGTACCCGGCGAAGGTAAAGTTCCGGGAGAGGGAAAAGTTCCCGGTGAAGGTAAGGTCCCAGGCGAAGGAA AAGTTCCTGGGGAAGGAAAAGTGCCCGGGGAAGGGAAGGTACCTGGAGAGGGGAAGTGCCAGGAGAAGGCAAAGTACCTGGGGAGGGTAAAGTACCAGGAGAGGGTAAGGTT CCAGGTGAAGGGAAAGTACCAGGCGAGGGTAAAGTCCCCGGAGAAGGAAAGGTACCCGGAGAGGGCAAGGTACCGGGGGAAGGCAAGGTTCCTGGTGAGGGAAAGGTCCCTGG TGAAGGAAAAGTCCCGGGTGAAGGTAAAGTACCTGGCGAAGGGAAAGTTCCT
[GGMGSP]28: GGTGGAATGGGCTCCCCTGGAGGGATGGGAAGTCCCGGAGGCATGGGTTCTCCTGGCGGTATGGGGTCACCAGGAGGTATGGGTTCCCCGGGAGGAATGGGGTCCCCAGGTGGTAT GGGTAGTCCCGGCGGAATGGGTAGTCCAGGGGGTATGGGAAGTCCAGGAGGAATGGGTAGCCCAGGAGGGATGGGCTCGCCTGGTGGAATGGGAAGCCCTGGTGGGATGGGTAGT CCGGGAGGTATGGGATCTCCAGGTGGAATGGGTTCACCCGGAGGAATGGGATCTCCTGGAGGAATGGGCAGCCCCGGAGGTATGGGCAGTCCTGGTGGTATGGGATCGCCAGGGG GAATGGGATCACCGGGTGGTATGGGCTCACCTGGTGGCATGGGTAGTCCTGGAGGTATGGGGTCGCCTGGGGGCATGGGATCTCCGGGCGGCATGGGCTCTCCAGGCGGGATGGGA TCTCCCGGTGGTATGGGGAGTCCTGGGGGTATGGGTTCGCCC
[GGHILP]28: GGTGGGCATATACTTCCAGGGGGACATATTCTACCCGGAGGTCACATTCTTCCTGGCGGTCATATCTTACCAGGAGGTCATATTTTGCCTGGGGGACACATCCTTCCAGGTGGTCACA TATTGCCGGGTGGACATATCTTGCCAGGTGGACATATTTTACCAGGTGGCCATATCCTCCCCGGCGGCCATATATTACCAGGCGGGCATATTCTCCCAGGAGGACACATACTTCCCG GGGGTCATATTCTGCCAGGAGGCCATATACTACCAGGGGGTCACATCTTACCTGGAGGGCATATCCTTCCTGGTGGTCATATACTTCCTGGAGGTCATATCCTGCCTGGAGGACATAT ACTTCCGGGAGGTCATATATTACCGGGCGGACATATCCTACCTGGTGGACATATATTACCCGGTGGACACATTTTACCTGGGGGCCATATTCTTCCAGGAGGGCACATTCTCCCTGGA GGCCACATACTCCCAGGTGGGCACATATTACCT
[GGSHVP]28: GGTGGCTCTCATGTGCCTGGAGGATCGCATGTCCCTGGGGGTTCACACGTTCCGGGCGGTAGTCATGTACCAGGAGGATCTCATGTCCCCGGAGGTTCGCATGTGCCCGGGGGCTCG CATGTACCCGGAGGAAGTCACGTACCTGGTGGTTCCCATGTACCGGGAGGTAGCCACGTTCCTGGCGGATCACATGTACCTGGAGGTTCCCACGTACCCGGCGGGTCGCATGTTCCG GGAGGCAGCCATGTTCCCGGTGGAAGCCATGTGCCGGGAGGATCACACGTACCAGGGGGATCGCACGTTCCCGGAGGGTCCCACGTCCCTGGAGGGTCACATGTCCCAGGTGGTA GTCACGTGCCAGGCGGCTCACACGTCCCAGGAGGTTCACATGTTCCAGGAGGGAGCCATGTCCCGGGTGGCAGTCACGTTCCAGGTGGGTCACACGTGCCTGGTGGGAGTCATGTT CCTGGAGGCTCACATGTGCCAGGTGGATCTCACGTCCCC
[GSMGSP]28: GGTAGTATGGGTTCGCCTGGATCAATGGGCAGTCCTGGCTCGATGGGATCTCCCGGAAGCATGGGGTCACCAGGCAGTATGGGTAGTCCCGGGTCAATGGGTTCTCCTGGTAGCATG GGTTCCCCAGGGTCCATGGGTAGTCCTGGATCTATGGGAAGTCCTGGTTCGATGGGTAGTCCAGGTTCTATGGGTAGCCCCGGCTCAATGGGTAGTCCGGGTAGTATGGGCAGCCCA

GGCTCCATGGGATCTCCAGGTAGTATGGGGAGTCCAGGGAGCATGGGAAGTCCAGGATCGATGGGGTCCCCTGGCAGCATGGGCTCACCGGGATCTATGGGTTCACCCGGTTCAAT GGGATCTCCGGGGTCTATGGGGTCGCCCGGATCTATGGGCTCTCCTGGAAGTATGGGATCGCCGGGCTCTATGGGATCTCCTGGGAGTATGGGATCACCTGGGTCGATGGGCTCGCC AGGTTCCATGGGGAGCCCTGGTAGTATGGGAAGCCCT
[GTSGTP]28: GGTACAAGTGGTACGCCCGGTACTTCAGGGACTCCAGGAACTAGCGGAACCCCAGGCACATCTGGCACACCAGGAACGTCTGGTACTCCAGGGACAAGTGGGACTCCTGGGACTA GTGGGACCCCGGGCACAAGTGGAACTCCCGGCACTAGTGGCACCCCAGGAACCAGTGGTACACCAGGGACCAGCGGGACACCTGGTACATCGGGAACACCAGGTACATCAGGAA CTCCAGGTACTAGTGGAACCCCTGGAACCTCAGGTACACCTGGGACATCAGGCACTCCTGGAACAAGCGGCACGCCAGGTACGTCAGGTACTCCTGGTACGAGTGGAACACCTGG AACTAGTGGTACCCCCGGAACATCAGGTACCCCTGGTACTTCTGGGACGCCTGGTACCTCTGGAACGCCGGGGACTTCCGGTACACCCGGGACCTCAGGAACACCGGGAACGAGCG GTACTCCGGGTACATCCGGAACTCCTGGCACTTCGGGTACTCCC
[GYSGVP]28: GGTTATTCTGGAGTTCCTGGGTATTCAGGCGTGCCCGGATATAGTGGTGTCCCTGGCTACAGCGGTGTACCAGGGTACTCTGGTGTACCTGGGTACAGTGGAGTACCGGGATATTCTG GTGTGCCAGGTTATAGCGGAGTACCAGGCTACTCTGGCGTACCTGGTTATTCCGGCGTTCCTGGCTATTCCGGAGTTCCGGGATACTCTGGAGTGCCGGGTTATTCAGGTGTACCCGG ATACTCGGGGGTACCGGGGTATAGTGGAGTCCCAGGCTATAGTGGGGTCCCGGGCTATTCGGGAGTACCCGGTTATAGTGGCGTCCCCGGGTATAGCGGGGTTCCTGGTTACTCTGG GGTGCCTGGATATTCGGGCGTACCAGGATACTCCGGTGTTCCCGGCTATTCAGGAGTACCTGGATACTCAGGTGTTCCAGGATATTCAGGGGTACCAGGTTACTCAGGAGTTCCAGG GTATTCGGGTGTACCGGGTTACAGTGGTGTTCCT
[GYVGTP]28: GGTTATGTGGGCACGCCTGGCTATGTCGGTACTCCCGGATATGTAGGAACCCCTGGATACGTTGGTACACCAGGCTATGTAGGCACACCAGGATACGTGGGAACTCCAGGATATGTT GGAACTCCTGGATATGTGGGTACACCGGGATATGTCGGCACTCCTGGGTATGTGGGAACGCCCGGGTATGTAGGTACGCCAGGTTATGTAGGGACTCCAGGTTACGTGGGTACTCCT GGCTACGTAGGAACTCCCGGTTATGTCGGAACTCCGGGCTATGTGGGGACACCTGGTTATGTTGGCACCCCAGGGTACGTTGGGACTCCGGGGTATGTCGGGACTCCTGGTTACGTA GGGACGCCGGGTTACGTCGGAACACCAGGGTATGTTGGGACCCCGGGATACGTAGGCACTCCAGGCTACGTTGGAACACCCGGATACGTCGGTACACCTGGGTACGTAGGTACCCC CGGTTACGTTGGCACTCCCGGCTATGTTGGTACTCCA
[GANMPPGASIPPGANIPPGASIPP]7: GGTGCCAACATGCCTCCAGGAGCGAGCATACCGCCAGGGGCTAACATTCCTCCGGGTGCAAGCATCCCACCTGGAGCTAATATGCCACCGGGGGCGTCCATACCCCCTGGAGCAA ACATACCACCAGGAGCTTCGATTCCGCCAGGTGCGAATATGCCCCCAGGAGCAAGTATCCCTCCTGGCGCTAATATTCCACCGGGCGCTTCAATCCCCCCTGGGGCCAATATGCCTC CCGGAGCTAGTATTCCCCCTGGTGCTAATATACCACCGGGAGCCTCAATTCCGCCCGGAGCAAATATGCCGCCTGGAGCCAGCATTCCCCCAGGCGCAAATATTCCGCCTGGTGCAT CAATTCCCCCGGGAGCTAACATGCCACCCGGTGCCTCGATACCACCCGGGGCAAATATACCCCCAGGTGCTTCCATTCCACCCGGCGCGAACATGCCCCCCGGAGCGTCAATTCCA CCAGGTGCAAATATCCCGCCGGGAGCATCTATACCTCCT
[GVPNPPGVPKPPGVPKPPGVPNPP]7: GGTGTCCCAAATCCTCCCGGAGTACCAAAGCCTCCAGGGGTTCCAAAACCGCCAGGAGTGCCGAATCCACCTGGGGTACCGAATCCTCCTGGTGTACCTAAGCCGCCGGGAGTACC TAAACCACCCGGTGTACCCAATCCTCCAGGAGTACCCAACCCACCTGGCGTTCCGAAGCCACCAGGTGTACCGAAACCTCCAGGCGTACCTAATCCCCCGGGTGTACCAAACCCTC CCGGGGTGCCCAAACCACCGGGCGTACCCAAACCTCCTGGGGTCCCTAATCCACCAGGCGTTCCTAACCCGCCCGGAGTCCCGAAACCACCAGGGGTACCAAAACCTCCGGGTGTT CCAAATCCCCCTGGGGTTCCTAATCCGCCTGGCGTCCCTAAACCTCCCGGTGTTCCCAAGCCCCCCGGTGTGCCTAATCCTCCGGGGGTACCTAACCCCCCAGGCGTGCCAAAACCA CCTGGAGTTCCTAAACCCCCTGGCGTACCAAATCCACCC
[GIVLPPGIVLPPGIVFPPGIVLPP]7: GGTATAGTCTTACCGCCCGGAATCGTTTTACCCCCTGGAATTGTTTTTCCTCCTGGGATAGTTTTGCCACCTGGTATAGTATTGCCTCCAGGTATCGTGTTACCTCCCGGCATAGTATTC CCACCAGGAATCGTATTACCACCAGGCATCGTACTACCTCCGGGTATTGTATTACCTCCTGGTATTGTCTTTCCACCCGGTATTGTTCTACCACCTGGAATCGTCCTACCTCCAGGCAT AGTGCTTCCACCAGGTATTGTGTTTCCACCGGGAATTGTGCTCCCTCCAGGAATAGTTCTGCCTCCTGGAATAGTACTTCCCCCAGGAATTGTATTTCCTCCAGGGATAGTCCTGCCAC CAGGGATCGTTCTTCCGCCTGGTATCGTCTTGCCCCCCGGGATAGTATTTCCACCTGGGATTGTACTCCCGCCGGGGATAGTGTTGCCGCCAGGTATAGTTCTCCCACCTGGCATAGTT TTCCCTCCTGGCATTGTCCTTCCTCCC
[GAQGSP]28: GGTGCCCAAGGTAGTCCAGGCGCTCAGGGTTCTCCTGGGGCTCAAGGAAGTCCCGGAGCACAAGGGTCACCAGGTGCGCAGGGATCTCCTGGCGCACAGGGGTCACCCGGTGCGC AAGGGTCCCCCGGCGCCCAAGGCTCCCCAGGTGCCCAGGGTAGTCCTGGTGCACAAGGCAGCCCAGGAGCGCAAGGTTCTCCCGGGGCTCAGGGGAGTCCAGGGGCACAGGGCTC TCCGGGCGCGCAAGGATCACCTGGAGCCCAAGGATCTCCAGGAGCCCAGGGATCACCAGGCGCACAAGGAAGCCCTGGAGCGCAGGGTTCACCAGGGGCGCAAGGCTCACCAGG AGCACAGGGTAGCCCGGGTGCTCAAGGTTCCCCTGGCGCTCAAGGGAGTCCTGGAGCTCAGGGATCGCCTGGGGCACAAGGTTCACCGGGAGCTCAAGGCTCGCCAGGTGCTCAG GGCAGTCCAGGTGCACAGGGAAGTCCGGGGGCCCAAGGGTCGCCC
[GSFGIP]28 (NR): ATGAGCAAAGGACCAGGTTCATTTGGCATTCCGATAGGTGGGTTCCCTATTGGATTTTTCCCAGGTTTCAGCGGTGGAAGTATACCAGGTGGTATTATATCACCAGGGGGCAGTGGG AGTCCTGGTTCGCCCGGCATCGGGGGAGGAATCCCTGGGGGTGGTTTCGGTTTCTTTCCTATTAGTGGTGGCTTTCCGGGAATAGGTTCTCCATCTATAGGATTTCCAATCTCAAGTGG AGGATTCTCTTTTCCAGGATTTGGAATTGGTTTTCCTGGCGGATTTATTGGTAGCCCAGGAGGTATTCCTATCATAGGAGGCTCTAGTTTTATACCGGGTATACCTATATTTGGTATCTT TGGAGGGTCACCTTCAAGCTCAGGTAGTCCGTTTGGATCAATTTCTATACCTGGATTCGGCTTCCCTTCTGGAGGTTTTGGGATAAGTCCAGGCATACCTTTCTTCTTTATTATTCCAGG GTTTCCTTCGTTCGGAGGTGGATTCATACCATCAGGATCTCCTGGTTGG
[GDGYAP]28 (NR): ATGAGCAAAGGACCAGGTGATGGTTATGCCCCTGCTGGAGCTTACGCGGCCCCAGCTGCTGCCGATGGTGGATATGGAGATCCTGGAGGTGCGCCAGGAGATCCAGCGGGAGGTTA CTATCCATATGATGATGCGTACGGTTATTATCCCGGCGCTCCGGCTGGTGGCGGAGATGATCCTTACGGGGCTCCTGGTGCTTATTACCCTTATGGCTATGGAGGATATGCACCTGGT GACGCCCCCGCAGGTTATGGATACGATCCAGGAGGTGACCCATACGGAGGTGGTGCAGGCGGTGCTCCTTATGATGGAGGGGATGGTGATGCTCCCTATGGGGGAGACGGTGGAG GCGACCCTGGTTACGATGGCGATCCGGGTGCTGATGCAGACCCAGGTGATCCTGATGGTGCCTACGACGCTCCAGGCTATCCAGGTGGGTATCCTTATGGTGGTTATCCTGGGGACC CCGGTGATTATCCAGGGGGTGGAGATGGAGCAGGAGCTGATCCAGATGCTTACGGTTGG
[GYGHAP]28 (NR): ATGAGCAAAGGACCAGGTTATGGACATGCTCCGCATTACGGTGGAGGTCATTACTACGCGCCAGGTTATCCACATGGAGGACATCCAGGAGGCGGTGCACATGCACCAGGCCATCC TTACCATTATGGAGCAGCAGGAGCACCTTATGCAGGCTATCCAGGCGCTCCTTATCATGGGTATCCAGGTGGTGCTGGTGGTCATGGTGGAGCGGCTGGGGGTGCCCCAGGAGCTTA TCCAGGGGGACATCATCCTCACGGAGGTGCAGGTCACGCCCCCGGCGCAGGACACGGGGCTCCCCATGGACACCCAGGTCATGCATATCCTGGAGCTGGAGGGGCCGGAGCTCAC TACCCTTATGGTCATCCGGGTGCTTACGCACCTGGTCACTATCCTGCCTACGGACATGGTTACCCTGCTCATTATTATCATCACCCTGGGCATCCTGGTTATTATCCTGGCTATGGATA TCCCGGTCACCATCCCTATGGTGCCGGTGGGTACTATCCAGCGGGAGGTGGCGGTTGG
[GGYSAP28]28 (NR): ATGAGCAAAGGACCAGGTGGGTATAGTGCCCCTGGTAGCCCTGGAGCCGGAGGGGGTGCTCCAGCCTATGGTGGCGGTGCCGCTGCTCCTAGTTCTGGTTCAAGTCCCTACGCTGGT GGTAGTTACCCTGGGGGAGGAAGTGGATCAGCTGCCTCACCAGGTAGTGCATCTTCACCTTATGGAGGGTATCCTTATTCTTATGCGAGTGGTGGAGGATCTGCACCTGGAGGTGGT TATCCTGCTGGCTATCCATCGGGTTATCCCGGTGGTGCGCCAGGTTCGTATTATAGTGGAGCGCCTGGTTATGGGTCAGGAGGCGGAGCATATCCTGGCAGTTATCCAGCTTCTGCTT ACTATCCAGGAGCTCCGGGTGGGGCCTACAGTCCAGGCTATTACGGAGGTTACTACCCAGGTGGATACGGGGCGCCGGGGGCTCCTTACTCTGGAAGCCCAGGAGGATATCCAAGT TCATATGGAGCTCCCGCAGGAGGTGCTAGTCCTGGTGCTGCAGGTTATCCGGGTTGG
[GYGYAP28]28 (NR): ATGAGCAAAGGACCAGGTTATGGGTACGCTCCTGGTGGAGGCGCTCCAGCGGGAGGTGGATATTATTATCCTGGAGGATACTATCCAGGTGCGCCAGGAGGTGCTTACGCCCCTGG TTACCCAGGTGGTGCACCTTACTATGGTGGTGGGTATGGAGGTTATTATGGTTATTACCCCTATTATTACGGTGGAGCAGGCGGTGCCCCCGGTGCAGCACCAGGTTACTACGCGCCG GGTGCTCCTTATGGATACGGGGCCTATGCTTACCCTGGAGCTGGTGGCTATGGATATGGAGGGGGAGCATATCCTGGGGGTGGTTATCCGGGAGCTTATCCCTACGGTTATCCTTACG GAGGGTATCCTTATTATGCCTACTACTACCCTTATGCAGGTTATCCAGCATATGGTGCCTATCCTGCGGGTTACGGCTATTATGGAGCCTATTATCCATATGGGGCTCCGGGCTACGG ATATCCAGGCTATCCATACGCATACTATGCACCCTATGGCGGAGCTCCTGGTTGG
[GVGYAP28]28 (NR): ATGAGCAAAGGACCAGGTGTAGGCTATGCGCCAGCCGTCGCTGCTGCACCTGGTGTCGTGCCAGGTGGAGGATATGGTGTTCCATACGGTGGATATTACTATCCATATTATTACGCA GCCCCAGGGGGAGGGGTTGTCTATTATCCCGCTGGTGTGCCTGCTGGAGGGTATCCAGTAGGTGCCCCCGGCGGATATGCTCCTGGCGGTGTACCTTATGTAGGAGCCGCACCGGGC GCTCCAGGAGTCGCAGGGGCCGGAGCTGGGGGTGTTGGTTACGGAGGTGGCTATCCTGGTGCTGTACCGGCAGGTTACCCTGGAGTGGGTGGTGGTTATGCAGGAGTTGGAGCTTAT GCTTACCCAGGAGGTGCTCCCGGTTATCCAGCTGCCTATGGAGGCTACGTCCCTGGTGGGGTAGCACCCGGAGTTTATGGTGGAGCAGTTCCTTATGGAGTACCTGGAGGTTATTATC CTGTGGGAGTAGGAGTAGCTGTCCCAGGTGTATATGTACCAGGCGTTCCCGGTTGG
[GHGNRP28]28 (NR): ATGAGCAAAGGACCAGGTCATGGAAACAGACCCGGAGGTGGTCATCCTGGTAATGGACATCCTCATGGTAGGCCAGGGAGAGGAAACGGAAGACCAGGTAATCACGGTGGTGGC AGGAACGGTCATCATCCCGGTAGACCTCATAGAAGACATAGACACGGAGGAAATCATCGTAATCCGCATGGAGGCAATCCTAATGGTGGAAATGGTCGGCATCCAGGAAATAATC GCCCTGGAAATCCTGGCAACCCAAGGGGTAACCCTAGACATCGACACAGACATCATCACCCAGGGGGTCATAGGAATCCTCGTGGACACCCCGGCCATCATGGAAGAAATCCAGG TGGGAATCCACATGGTCACAATCCAGGCAGACCTGGGCACCCTGGTGGTAATAATCCTGGGAACAATGGGGGACATGGAAACCATCCGGGTGGAGGTAATAGACGTAGGGGAGGT CGTCCTGGAGGACATAATCCAAGAGGTAGAAGGCGTCCAGGAGGGAATGGCGGTAATCCCGGTTGG

[GGHNIP28]28 (NR):
ATGAGCAAAGGACCAGGTGGGCATAATATCCCTGGTATAGGGCACCCTGGAAACGGTGGTATACCTATAGGAGGTGGTCATCCGGGTATTATTCCAGGTAATAATCCTCATAATGG
ACACGGTCATAACAATATACACCCCGGCATCCCAATTAACATACATCATCATCCAGGTGGAGGTATCCCAGGTCATGGAGGAATTATCAATGGAGGGAATGGTGGTGGAAATCCAA
TAGGAAATATTCACAATCCAGGAGGACATGGTAATCCCGGTATTGGTAACCCAGGGGGTATAAATGGTATTCCTGGCATTCCAAATCATGGGGGAGGTCACATACCAGGCCATAAT
CCAGGCATACCTCATCACGGAGGCAACGGAAACCATCCTATCGGACATCCACACCATGGAATCCCTAATGGGAACAACATTCCTAATCATATACATATTCCGCATGGCGGTAATAT
ACCTGGTGGCAATCCTGGTAATGGTCACCCACATGGAAATGGAATTCATCCTGGGGGTTGG

[GLNGVP28]28 (NR):
ATGAGCAAAGGACCAGGTCTTAATGGAGTGCCTGTAGGTGTCCCTGGACTTCTTCCAGGCGTTCCAGTCGTAGTTAATCTACCAAATGGTGGGGTTCCGGGCAATGGAAATCTTGGA
GGCAACGTAAATAATCCTAAACGGTGGAGGGAATCCTGTGAATTTAAACTTACCTTTGTTATTAAATTTGCCGGGTGGTGTACCAGTACTTGGTTTACCAGGTGGCCTTCCTGGTGTTG
GTGGTAATTTACCTGTTCTCGGGGGCGGTGTAGTACCGGGAGTAGGGCTAGGAGTTCCTTTAGGAGGAAACAATGGGAATGGAGGACTATTACCCGTAGGAAATGGCTTAGGTAAT
CCAGTTGGAGTCAATGTTCCTGGAGGTGTGCTACCTGGGGGAGGTAATGGTAACCCTGGTCTAGGTGGAAACCCAGGGGTCCCAGGAGGGGGTGTATTACTTCCAAACGGAGTTGT
CAATCCTGGCGGAGGTTTGAATCCAGGTCTGCCAGGTTTAGTAAATCCGAATGGTTGG

[GYGNTP28]28 (NR):
ATGAGCAAAGGACCAGGTTATGGAAACACGCCAGGAACACCTTATTACACAGGAAATTATGGGGGCACTTATACACCAGGCGGTAATCCTACAGGGAACCCCGGTGGTGGCTACC
CAACAGGCACACCTGGTTACAATACCTATCCTGGTACTGGAAATGGAAATCCAGGTTACGGAAACGGTTATACGACACCCTACGGTACGGGCTATCCGGGTGGTAATGGAGGCTAT
GGTAATCCCTATGGAGGTAATAATTATCCTTACGGCGGAACTGGTAACACTCCTTATGGCAATCCAACGTATGGAACGAATCCTGGAGGATACCCTTATACTAATTACTATCCCGGC
AACCCAGGTAACCCGACTACTAATACAAATCCTAATGGGAATCCGAATGGAACTCCAAATACTTATAATGGTACATATGGTTATCCTACCAATGGATATGGTGGAGGGACCCCAGG
AGGTGGGACAGGTACTCCTAACGGAGGTTATTATCCAGGGGGAGGAAACCCTGGTGGTTGG

[GGHQQP28]28 (NR):
ATGAGCAAAGGACCAGGAGGACATCAGCAACCTGGAGGGCATCAACCCCAAGGTCACCCACAAGGTGGAGGTCAAGGTCAGCATCCCCACGGTCAACAACCTCAAGGACAGGGG
CAGGGAGGTGGGCACCCTGGTCATCCAGGAGGCCACCCTCAACAACAAGGTCAACAGCACCATCCTGGCCAAGGCGGCCAGCAACAACCAGGACAACAGCCCGGTCATGGCGGA
CAACACCAACCAGGCCATCAGCCTGGACAAGGACAACCGGGTCAACACCCCGGACAGCAGCAAGGTCATCATCCTGGGGGACAACCTGGTGGTGGACAACATCACCAGCCAGGT
GGTCAACATCCTCAGGGCCAACCTCATGGACAGCCGGGACATCCACACGGCGGGCAGCCTCACCAACAAGGCCAACATGGGCAACCGCAACATCAACAAGGGGGTCAGCCACAA
CAGGGACACCCAGGTCAACCACAGGGCGGTCAGGGTCATCAAGGGCAAGGACATGGTCAACCTGGTTGG

[GGQHIP28]28 (NR):
ATGAGCAAAGGACCAGGTGGGCAGCACATTCCTCATGGTATTCCTGGTCAACATCCGGGCCACGGTGGTCAGCCGATTGGAATCCCTGGACATCCTCACGGAGGTATACCAATAGG
AGGACACCCCGGTATTCATCATCCCGGTCATATACCCGGCATACCTATACATGGAGGTCACCATCCTGGCGGCCAACCAGGTCAGATACAGGGGCACCCAGGGCAACATCATCAGC
AGCCTGGTATCGGACAACCACAGATTATAGGACAAATTCCAGGCATCCAACAGCAACCTGGGGGACATGGGATCATTCAACCGCATCAAATAGGTCAACCCATACAACACCAAGG
AGGCCATGGTCATCCAGGAGGTGGACAACAAGGTCATGGAATACACGGCGGAATACATATCCCAGGACATCAACAACCTCAGGGAGGGATACCTCAAGGACAGCCAGGTGGTGG
CATTCCACATGGTGGAGGAATTCACCCTGGTGGTATACAAATACCTGGAGGGGGTCATGGTTGG

[GVMGQP28]28 (NR):
ATGAGCAAAGGACCAGGAGTTATGGGCCAGCCAGGTGTAATGGGGCAACCAGGAATGGGTCAAGTACCTGTCGGTATGCCAGTCGTTATGGGTGGCGGAGTAGTACCCGGTGGAG
GCATGGGTATGGGGATGCCTGGGATGCAAGTTCCTCAAGGGATGGGACAAGGTGTGCAACAAGGCGTTCCTGTACAAGGAGTCCAAGGTCAACCAGTAGGAGTTCCAGGTATGCA
GCCTCAAATGGGAATGCCTATGGGAGTACCAGGGGGAGTGCCAATGCAGGTTATGGTTCCAGTCATGGGCATGCCACAGGGAGGTATGATGCCAGGAGGACAACCGCAAGGAGGT
CAGCCTGGTGGAATGCCAGTTCAAGGTGGGGTAATGCCTGGAGGGGTTCAACAACCTGGTCAGGGCGGGGGTGTTCCACAAGTTGGTGTACCAATGATGGGTATGCCTGTTGGAGG
AATGCAAATGCCCGGAGGTGGACAGCAACCTCAGCAGGGTGTCATGCAACCCGTAGGTGGTGGTATGCCGGGAGGAGGTTGG

[GVEGIP28]28 (NR):
ATGAGCAAAGGACCAGGTGTTGAAGGCATTCCGGGAATCCCAGGAGTGGGTGAAGGAGTCGAAGAAGTAGAGGAACCTGAAATTGGAGTACCTGGCATCATACCTGGAGTTCCAG
TTGGAGAGCCAGGAGGTATTGAGATAGAAATTCCAGGGGGTGTACCTGTGATTGGTATCCCTGGTGGTATTCCAGGCGAGGGTGGAATCATTCCTGTCGGAGAAGGTATAGGAGAA
CCCGTGATAATTGAAGAACCAGAAGGAGGCGTACCAGGAGAGGGAGAAATACCGGGTATAGTTGTAGAACCTGGGGGAGGAATCGAACCAGGAATTGTTCCTGAGGGCGGTGAG
GTCATCCCCGGCATAGGAATAGGTGTCCCTATTGGTGGAGAAGAAGGGGAAGGTGGTGAGCCTGGTGTGCCTGGCGGAGGGGTCGTTCCAGGTGTAGGTATCGGTGTAATAATAGT
AGTTCCTGGAGGTGAACCAGTAGGAGTAGAAGTGCCCGGTGGGGTTATACCAGGTGGCGGTTGG

[GEGKVP28]28 (NR):
ATGAGCAAAGGACCAGGTGAGGGTAAAGTTCCAGGAGTCGAACCAGTTGGAGTAGTAAAAGTGAAAGTCCCTGGAGGAAAAGGGGAACCTGTAGGAGGCGTACCAGGGGGAGTG
CCTAAGGAAGGTAAGGGTGGGGGCAAAGAGGAACCAGAGGGCGGCGGTGAACCTGAGGGAGTACCTAAGAAAGGAGAACCCGGAGGTGGTAAACCTGGAGAAGTTGAAGAACC
GGAAGAAGAAGGTGAAGGAGGTGAGCCTGGAAAAGTAGAAGAGCCAGGAGGGGTTGTTCCTGGTGTTGGGAAACCCAAGGGAGGTAAAGAACCTGGGGTCCCAGGAAAGCCCG
GTAAAAAACCTGGCAAGCCTGTTAAAAAAGTCGTTCCTAAAGGTGTACCCGAAGGTGGAAAGAAACCTGTCAAAGGCGGGGGTGTCGGTGGTGAAAAAGAAGGAAAACCAGGTG
GCGGAGTTAAGAAGCCAGAAGGAGTTGTAGAACCAGGCGAGCCGGGAGAGGTTGGTGGAGGTGTGGGTTGG

[GGMGSP28]28 (NR):
ATGAGCAAAGGACCAGGTGGAATGGGCTCCCCTGGTTCAGGGGGAGGTTCTATGCCTGGCGGAGGGTCGGGTATGGGCGGTGGTAGCAGTCCAGGAGGTGGGATGCCTTCTGGTGG
AGGATCAGGAGGAAGTGGAGGCGGCATGAGTCCGGGAGGTAGTCCAATGGGTGGTGGGTCTGGAAGTCCTGGTGGTATGGGGATGCCAGGTGGTTCTCCTGGGATGGGAGGCTCTC
CGTCAGGTGGCGGTATGCCTATGGGTATGCCAGGAATGCCCTCAGGATCTCCAGGTTCACCTGGGGGTGGCAGCCCAGGGTCTAGCCCTGGAATGCCTGGGTCACCAGGTATGCCG
ATGGGTGGGGGAATGCCAGGCGGGGGTATGATGCCTGGAGGTATGGGTGGAAGCGGAATGCCGGGGTCCATGCCTGGTATGCCCGGAGGCATGCCAGGGAGTGGTGGCTCACCCG
GTGGATCTGGAGGGATGCCCGGTATGGGAGGGGGCATGGGAGGAATGGGAGGTGGAGGTTGG

[GGHILP28]28 (NR):
ATGAGCAAAGGACCAGGTGGGCATATACTTCCAGGCATTCCTGGGGGACATCCTATCGGTGGTCACGGAGGATTGCATCCAGGAGGTATTTTAATAGGAATCGGACATGGTGGAGG
TTTAATACTACCTTTGTTGGGTCATATTATTCCCATAGGACACCCTGGAGGCTTGCCAGGACTAGGTTTGGGAGGTCTTCCTCTTGGACATCATCCTCACATTCTAATTCATCTGCCGG
GCCATCCTGGTCTGATACCTCATGGGTTACCAATCTTAGGTGGATTAGGGGGCGGAATCATTGGTCATGGAATCCCAGGTATAATAATTCCAGGTGGCTTACTCCCTGGTTTACCGCA
TGGTATTCCTTTAGGATTACCTATACACTTACATCACCCAGGGCTACCAGGTCATTTGCTTCCAATAGGTCACCATCCCGGTATAGGAGGAATTGGAGGACATATTCCACATGGAGG
GATACCACACGGTCATCCGGGTATACATCCATTACACCACGGGGGTGGTGGTTGG

[GGSHVP28]28 (NR):
ATGAGCAAAGGACCAGGTGGCTCTCATGTGCCTGGGGGTGGGTCGCACCCAGGAGGTCATGGAGGTGGTTCAGGTGTCGTACCAGTTAGTCACTCAGTAGTACATCATGTACCTGGT
GTGGGATCACACGTACATGGACACAGTCCTAGTGGTAGTCCTGGCTCGCCCGGTGTAGTGCCGGGTGTTCATGTTCCTGGTCACGTTCATCCAGGTCACCATCCATCAGGAGTACCG
GGAGGATCTGTAGGAGGAAGTGGATCTCCAGTAGGTAGCGGAGGTAGTGTACATCCACATTCGGGTCATTCACATCCGGGGCATCCTGGAGGCTCACCTGTACATTCTTCTGTCCCA
GGGGGAGTTGGACATAGCAGCCCTGGTTCTAGCCCAGGTGGAGGACATGGGGTCCCTGGCGTGCCAGGTTCTGGATCAGTTGTACCAGGCCATGGAAGTCCAGGCGGTCATCACCC
TGGAGTCGGTGGTGGTCATCCTCACGGACATCCTGTTGGAGGGGGTTCCAGGTGGTTGG

[GSMGSP28]28 (NR):
ATGAGCAAAGGACCAGGTAGTATGGGTTCGCCTGGATCTAGTCCAGGATCACCTGGAATGGGAAGTGGATCAATGCCTGGGTCACCGATGGGTATGCCAGGTTCCAGTGGAATGCC
ATCTAGTGGTATGAGTCCTGGTTCTGGTAGCTCTGGAAGTTCTTCACCAGGTATGCCCTCAGGCTCTCCAGGGGGCGGATCAGGGATGTCAGGAAGTCCTTCATCTGGAGGGTCCATG
CCAATGGGTGGAAGCGGGGGAGGAATGTCTCCAAGTAGTTCTATGTCCCCAGGAGGTTCTCCGAGTGGTGGGGGTAGTAGCATGCCTTCGGGAATGAGTTCACCCGGAGGTATGTC
GATGATGCCCGGTGGCGGTATGCCGGGAGGCAGCAGTCCGTCTGGGATGCCAGGAATGATGCCTGGTATGCCTATGGGCTCGTCAAGTGGTTCAGGTAGTGGAGGATCTGGCTCAC
CTATGATGGGATCTCCATCCATGGGGTCTTCTCCTATGGGAGGTGGTAGTCCCGGTTGG

[GTSGTP28]28 (NR):
ATGAGCAAAGGACCAGGTACAAGTGGTACGCCCGGATCTACTACCCCTGGAAGTGGAGGGACACCTACAGGTGGTACAGGAGGAACGACTGGTGGTGGAACAGGAAGCCCAGGT
ACGTCACCAACTACAACCGGAGGCTCTCCATCAGGAACAACATCTCCCACTGGAGGTTCAGGTACACCGACTACTGGATCAACTAGCCCTGGTACTGGTACAACTCCAACAGGCAC
GAGTCCAGGAGGTGGGTCAAGTCCTGGGACGACACCAGGAACTGGGGGTGGCTCACCTGGTTCGCCTGGAGGATCGACTCCTGGCACTCCTACAAGCGGAACTACTACACCTGGCG
GGGGCACACCTTCAACAACTGGAACCCCAGGGACCACAGGATCTCCTACTACGGGAGGTACTTCTACAACAACGGGTACATCTACTCCAGGTTCTTCTAGTCCCGGCGGTACTACTC
CGTCTACAGGGGGAACACCCGGTGGTTCTCCGACAGGTACTCCTTCTGGAACTCCCGGTTGG

[GYSGVP28]28 (NR):
ATGAGCAAAGGACCAGGTTATTCTGGAGTTCCTGGTGGAGGAAGCCCTGGTTATGTTTATAGTCCAGGTAGTAGTGTTTCTCCAGTTTATGGCTATGGTTCTTCACCTTACGGAGGTGT
CTCTCCTGTCGGATACGGTGGCGGAGTGTATGGAGGTTACGTCCCAGGTTATGGGGGAGTATATCCTTATTCAGTCTCAGTACCAGGTGGGTCACCAAGTTCATATGGTGGTAGTTAT
GGAGTACCTGTATATTCTTATGTCCCTGTTGGAGGCTATCCCTACGGGGGCGGTTACTCTTATCCAGGAGTCGTTGGTTACCCTGGAGGGGGTGTACCTTATGGTGTTCCAGGCTCACC
CTATGGATACCCAGGCGTACCAGTAGGATCTGTACCATCAGGAGGTAGCTCTTCGCCGGGTGGTGTAGGTGTGCCCGGTGTAAGTGGATATGGAAGTCCTGGCAGCCCAGGGGTTGT
CTATCCTGGGGTAGGAGGATACAGTCCGGGAGGTGGTTATCCTAGTGGTTGG

[GYVGTP28]28 (NR):
ATGAGCAAAGGACCAGGTTATGTGGGCACGCCTACAGGAGTGCCAGGAGTAGGTTATACTACTGTCTATGGTACTCCAGGTTATGGTTACCCAGGTACGCCGGGCGTTCCTGGTGGT
GTACCGACTGGCTATCCTGTTGGTACCCCCGGTGGTACAGGTGGAGGCTATGGATATGTTCCAACTTATACATATGGAGGTTACGTCCCAGGCACACCCGTTGGGTATGGAGTCCCT
GGAACTACTTATCCAGTAGGATATCCGACAGGGGGCTACACTGTAGTCACACCAGTGGGTGGGACACCTGGTACGACACCAACTGGAGGGGTTGGAGGATATGGTGTTTATCCTTA

CGGTGTTGGATACGGAGGTACTGTTCCCGGAGTTCCGGTAGGTGGCGGTTATCCTGGGGTCTACCCTGGTTATCCCTATGGGGGAGGTGTAGGTACATATACTCCCGTAGGAGTTTAC
TATCCAGGAGGAACACCTTATGGAACCACAGTTCCTACATACGTTGTAGGGGGTTGG

[PGGVVA]40:
ATGAGCAAAGGACCAGGACCAGTAGGAGTACCCGGCGTAGGGGGTGCACCTGTAGTAGTTGGGGGAGCTCCGGTTGGAGTGGGAGGTGCGGGTGGTGTAGTACCAGGCGCTCCAG
TTGGAGCTGTTGCAGGAGTCGGAGCACCTGGTGGTGCTGTACCAGCAGGCGGTGCTCCCGGAGTAGGAGCTCCAGGAGTTGTACCTGGCGCGGTACCAGGAGTAGGTGTAGGAGGG
GTAGGTGGTGGGGTTGTCGTACCTGTTGTTGTAGGTGTTGGTGGAGTTCCAGCGGGAGCCGTTCCAGTGGGTGGAGTAGCTCCTGTGGGAGTCCCTGTAGGTGCCGGTGGAGGGGCT
CCAGCCGGAGGTGTTCCTGTAGCTGTCGGTGTACCAGTCGTGGGCGGCGGGGTCCCTGGAGTTGGTGTCGTTGGAGTCGCCCCTGGTGCGCCCGGTGCAGGGGTGGGTGTGGTTGCT
GGGGCAGGAGCTGCTGCAGTTGCTCCAGTAGGGGTCGCTCCTGGGGTACCTGGAGTACCGGGGGCCGTGGTAGGAGTGCCAGGAGGCGTTGCACCAGCTGTAGTCGTAGGAGCAGT
TCCTGGTGTTGTTGGAGTAGTTCCCGTTGGTGCTCCTGTCGTCGCAGGAGGTGTAGCAGTGGCTCCTGTTGGCGGAGCGGTTCCAGGTGTCCCAGGTGGCGTGCCTGGAGGTGGAGCA
CCCGGGGCACCAGTTGTAGGCGTCCCCGTAGGTGTCCCTGGCTGGCCG

[PGGVVI]40:
ATGAGCAAAGGACCAGGACCAGTAGTAGTAGGAGTTGGAGGTGTAGTTGGGGTCCCCGGAGTCATCGGAGGTGGGGTTGGAGTACCTGGCGGTGTTCCCGGTGTAGTAATTCCTGG
TGGAGTCGTGGTTCCTGTTGTTGGTGGTATTGGTGTTGTACCTGTAGTTGTAGGTGTGGTACCTGGGGGTGTAGGGATCGGTATAGGAGGGATACCAGGAGGCGTGCCAGTCGGAGTT
CCTGGAGTAGTCGTACCAGGGGTTCCAGTAATCATAATTGTAGTAGTTCCAATCGTAGGAGTCCCAGTAGGTATCGTTGGTATCATACCCATTGGAGGAATTCCCATAGGAGTAGGG
GTGCCCGTAGGTGGTGGCGTACCTGGAGGCATTCCTGGGATTCCTGTAGGAGTGGGTGGAGGGGTCGTTCCTGGCATCCCTGGGGTAGGTGTTATAGGAATAGGGGGCGGGGTACC
AGGTGTCGGGGGAGTTATACCTGGTGTCCCGGGTGTTGGAGTTGTTCCTATTGGAGTGCCGGTTGTCGTCGGTGGTGTAGTCCCTGTGGGAGGTATACCAGTTGGTGTGCCTGGAATA
CCTATTATTATAGTGATAGGTGTACCAGGCGGAGTAGTACCGGGCGTCCCTGGAATTATAATAGTTCCGGGAGTACCCGGCGTTCCAGGTGGAATAATTGGAATCATTGGTATTATTG
TGATTGTCATTCCAGGAGTACCAGTGGTAGGAGGTGGCTGGCCG

[PGGVVL]40:
ATGAGCAAAGGACCAGGACCCGGAGGATTAGGCGGCGGTCTAGTTCCCGTTGTAGGTGTTCCTGGTGGAGGTGTATTGGGTGGTGTCGTGCCGGGTGTTGGTGTGGTAGGTGTCCCC
GTAGGAGTAGGCTTGCCTGGAGTGTTACCAGGAGTAGTCGTCCCAGGGGTACCAGTGGGCGGGGGAGTGCCAGGAGGACTTCTAGTACCCTTAGGATTGGGAGGTGGAGTAGTTCC
AGGGTTACCTGGCGGAGGTTTGGTTCCTGGATTAGTACTTCCTGGGGGCGTACCTGGGTTAGGTTTAGGACTACCAGGCGTGGGAGTTGGAGGCGTCGGCTTACCTCTTGGTGTAGTG
GTGCCCGGTGGATTACCTGTTGGCCTTCCAGGTGTAGTACCGGGAGTAGTAGTAGGAGTGGTTGTACCTGGTTTAGTACCAGGTGGTGGTCTTCCACTTGTTGGACTAGGTGTGCCTCT
CGTAGTAGTTGGTGGGGTCGGTGTAGGAGTTGTAGTAGGGGGTGTACCATTAGGAGGGGTTGGAGTCGTACCAGTAGGAGGTCTGCCTGTGGGTGTCTTACCAGTCGGGGTTCCATT
AGTATTACTATTAGTTGTTCCAGTTGGGGTAGGAGTCCCTGGTGTTGTACCCGTCGTTGGAGTTTTAGTTCCTCTAGTAGGTGGCGTTCCTGTCGGAGTTCCGGTTCTTTTACTTGGAGT
ACCTGTAGGTCTACCTGGAGGGCTACCTTTAGGCTGGCCG

[PGGVVK]40:
ATGAGCAAAGGACCAGGAGTACCAGTAGGAAAAGTTGGGAAGAAGCCTGGAAAGCCAGGAAAGGTTCCTGTTGGAAAACCTGTGGTTGGAGGCGTTCCGGTAGGAGTAGTTGGA
GTTGTACCTGGCGTACCAGTCGTCGGCGGAGGTAAACCCGTAGGAGTTCCAGTTGTGAAAGGGGTACCGAAGGGAGTACCTAAAGGAGGAAAAAAACCTAAGGGAGGCAAACCA
AAAGGTGTGGGTGTGCCGGTCGGGGGTGTTGTCGTACCTGTCGGAGTAGGTGTAGTCAAAGGTAAGCCCGGAGTCGGTGGAGGTGTCCCAGGAGGTAAGAAAGGAAAGAAAGTAG
TTCCTGGAGGGGTTCCAGTAGTAGGAGTGGTACCTGTTGTTGGGGGAGTTGGCAAGCCGGGTGTAGTTGTTCCAGGTGTTCCTAAAGTAGGGGTCCCTGTAGGTAAAGGTGGCGGCG
TGCCCGTGGGAGGTGGTGTCGTTCCCGTTGGTGGTAAACCAGTGGGCGGGGTGCCAGGCGGTGTTGGTGTACCAGGGGGCGTCCCTGGAGTGCCTGGTGGAGTAGTACCCAAGGGT
GGTGGGAAAGTTAAAGTGAAACCGGGAGTCCCCGGTGTAGTAAAACCAGTTGGTAAGGTCAAACCTGGGGTTGTAGGAGGGAAACCAGGTGGGGTAGGGAAGGGTAAAGTACCT
GGTGTAGGTGGAAAAGGAGTCAAGAAACCTGTAGTGAAGGTAGGTAAGGGCTGGCCG

[PGGVVF]40:
ATGAGCAAAGGACCAGGAGTAGTAGGAGGTGTTGTTGGCGGAGTTCCAGTCGGGGGTGTCCCAGGAGGCTTTTTTGGTTTTTTTCCGGGTGGAGGGTTTGGAGGATTCCCTGGGTTTC
CTGTAGTCGTTGGTGTAGGAGGTGGGGTAGTAGTACCTGTAGGTGTTGGAGGCGTTCCTGGAGTGGTTGTACCCGGTGTTCCTGTGTTCTTCTTTGGTGTGGTACCTGGAGGTTTTGGA
GTACCTGGTGGTTTTGTCTTTGTATTTTTCCAGGGGTACCAGGATTCTTTTTTGTACCTGGCGTCCCCGGAGTAGTGTTTCCAGTAGGATTTCCAGTTGTCGGTGGTGGCTTCCCAGTG
GGCGTGCCAGGGGGAGTGCCTGGAGTAGGTGGAGTATTTCCAGGTGTGCCGGTTGGAGTTGTAGTGGGAGTAGTTGGATTTTTTGTTCCAGGTTTCGTGGTAGGAGGGGTCGTAGTAC
CGGGAGGTGTAGTAGGGGTTGTAGTACCAGTATTCGTAGGAGTACCAGGGTTCGTCGGAGTCGGCTTTCCAGGAGTCCCTTTCGGTTTTCCTTTTGGGGTTCCATTCGGAGGATTTGG
AGTCTTTTTCCCATTTGGTGGATTTCCTGGTGTTGTTTTTCCTGGGGTGCCCGGCGTACCAGGTGGTGTCGTACCTGGGGGCGGGGTCCCTGTTGTGGGTGTACCAGGCGGCGGTGTAG
TTCCCGTTGTAGGTTTCTTTCCTGGCGGCTGGCCG

[PGGVVS]40:
ATGAGCAAAGGACCAGGACCCGTAGGAAGCCCCGGCTCTGTACCCGGTGTCGTACCTGTAAGTGGTGTAGTTGTCGGTGGAGGTAGCGTCCCAGGTTCTAGTCCTGTAGTAGGGGTT
CCAAGCGGCGTTCCAGTAGGATCACCTGTTGTAGGTGGCGGAGTTAGCCCAGGAGTGGGTGTACCTGGAGTTGTAGTACCGGGAGTTCCCGTCGGGGGCGGGGTAGGTGTGCCAGG
AGGTGGAGTCGTTGGAGGGGTACCAGTGGTCGTAGGAGTATCAGTTCCATCTGGTGTCTCTCCTGGTGGTGGGGTTGTTGTTCCTGTTGGCAGTGTTTCGTCAGGTGTCCCCGTTGTAT
CCGTGGTAGGAGTCTCAGTAGTACCTGGCTCGGTCCCTGTAGGCTCACCTTCTTCTGGAGTGCCTGGTGTGGTACCTGGTTCTCCAGGGGGAGTAGTTGGTTCATCAGTACCCTCAGG
ATCGGGAGGTGTATCTGTTCCGGTTGGTGTTGGTGGATCAGGAGTCCCTGGCGTACCAGGGTCTGGATCTGTAGTTCCTGTGGGAGTCAGTCCAGTTGGGGGTGGTAGTGGAAGTGG
AGTAGGTTCAGGTAGTCCAGGTGGTTCGGTTCCTGGAGGAAGTCCGGGTGTTGTACCAGTCGGCAGCCCTGGGGTCGGAGGCGTGCCCGGAGTAGTAGGAGGATCTGGGTCACCAG
GTGTTCCAGGCGGTGTAGGAGTTGGAGTACCAGGTGGCTGGCCG

[PGGVVT]40:
ATGAGCAAAGGACCAGGACCCGTAGGAACACCCGGCACTGTACCAGGGGGTGTAGTTCCTGTAGGTGTCACTCCTGTGGGTGGTGGAACAGGTACCCCTACAACTGTAACTGTAGT
AGGGGGAGTAGTAGTACCTGGGGTGCCAGGTGGAGGAACTCCTGGCGTGCCCGTGGGTACTCCAGTGGTTGGAGGGGTCGGTGTTGGAACTGGAACCCCTGGAGGTACAGTCCCAG
GAGGCACACCAGGTGTGGTACCAGGTACGCCCGGAGGCGTAGTACCAGGAACCACAGTTCCGACAGGTACTGGAACAACACCAGTTGTGGGAGTACCTACTGTTGGTGGGGTAAC
ACCTGGAGTCGGCGGCGTCCCTGTTGTAGTTGGGACAGGAGTACCGGGGGTCCCAGTAGGTGGCGGAGTAGGAGTTGTTGGGGTTGTTGTACCCGGAACGGTTCCCGTCGGAGTGCC
TGGAGTGGTAGGAGGGACCGTACCAGTCGGGACTCCTACAACCGGAGTTCCAGGAGTAGTCCCTACCGGTGTGACTCCCGGTGGAGTTGGTGTCCCTGGGGGCGGGGTAGTAGGAG
GTGTCGTTCCTGGAGTTACTGTTCCTGTTGGCACAGTCACTACAGGTGTACCAGTAGTAACAGTTGGAGTAGTTGTCCCCGTTGGTACTGTAGGCGGTGTAGTAGTGGGAGGTGGTAC
GACTCCTGGTGGTGTACCTGGAACGCCTGGTGTAGGAGGGGGCTGGCCG

[PGGVVY]40:
ATGAGCAAAGGACCAGGACCAGTAGTAGTAGGAGTTGTTCCTGGCGGTGTTGGATACGGCTATGGAGGTTATCCTGGAGGCGTTCCCGTCGGAGTACCAGGAGTAGTGGTACCTGG
TGTTCCAGTATATTACTATGTTGTGGTTCCATACGTAGGAGTCCCAGTAGGATATGTAGGTTATTATCCCTATGGTGGTTACCCTTATGGAGTTGGGGTTCCTGTAGGTGGAGGGGTAC
CAGGTGGGTATCCTGGTTATCCAGTTGGAGTCGGTGGTGGGGTCGTACCAGGCTACCCTGGAGTAGGTGTTTATGGTTATGGGGGCGGAGTGCCGGGTGTAGGGGGAGTGTATCCAG
GTGTCCCAGGAGTGGGAGTAGTACCATATGGAGTACCCGTTGTTGTTGGTGGAGTAGTTCCGGTTGGAGGATATGGTGTAGTACCTGTTGTCGTTGGTGTCGGGGGTGTTGTAGGAGT
ACCTGGAGTTTACGGAGGTGGTGTGGGCGTCCCTGGTGGCGTGCCTGGAGTGGTCTATCCAGGAGGTGTAGTTGTACCAGTCGTGGGTGGATATCCTGTCGGCGTACCTGGCTATCCT
TATTATTATGTCTACGGTGTGCCAGGCGGGGTAGTCCCCGGAGTTCCAGGTTACTATTATGTACCTGGGGTGCCCGGTGTACCGGGAGGGTATTATGGATACTATGGATATTATGTGT
ATGTTTACCCAGGGGTCCCTGTGGTAGGAGGGGGCTGGCCG

[PGGVAL]40:
ATGAGCAAAGGACCAGGAGGGGCAGTTGTAGTACCCGGTGGAGCGGTTGGTGTGCCTGCCGGAGCCGTGGGAGGGGTCGGTGGTGTCTTAGTACCAGCTGTAGGAGCTGTTCCTGG
TGGTGGGGTACCTGCAGTGGGTGTTCCAGTTGCTGGAGTCGTACCTGGAGGATTACCGGGAGGTGCTGGTGTTGGAGCTCTAGGTGCGGCTGCGCCGGGCGGAGTTCCTGTCGGAGC
TGCCCTTCCTGGAGTGGCCGGTGTCGCCCCCCGGAGGTGGTCTACCTGGTTTAGGCGCGGGAGCTGGACTAGGATTGGCTGGGGCCTTAGTCTTACCAGGACTTGGAGGCGTACCAGG
TGTCCCAGGTGGAGGTTTGCTTCCGGGTGGAGTACCTTTATTAGGTTTACCTGGCGTCCCCGCAGGGTTGCCTGGAGTTTTACCAGTAGGTCTGTTACTTCCCGGGGCTCCAGGGGTTC
CAGCGTTAGGTGTACCTCTAGGAGTCCCTGGTGTAGCACCAGGTCTTCCAGGTGCAGGAGGCTTGCCAGGAGTTGGTGCCCCTGGGTTATTACCTGTAGGTGGCGCTCCTGCTGCTCT
TGGTGGATTAGGACTACCAGGGGGTGGTGCATTGGGATTATTAGCTCCTGGGGGAGCATTAGTTCCAGGAGTAGGAGGTGTGGCTCCTGTTGCAGCGGGTGCTTTGTTACCAGGCGG
TGCTCCATTAGGAGTGCCAGCTTTACTCGCTTTAGGCTGGCCG

[PGGVAT]40:
ATGAGCAAAGGACCAGGAGTTGTCGTCGCTACCCCCGGTGCAGGGACACCTGGTGGAACCCCTACTGTCGCGGGTGGTGCAACTCCAGCCACTACTGGAGTCGTGGCAGGAGGGG
CGCCGGGAGGTGGAGCAGCACCTGGAGCTGGCGGCGGTGCTCCAGCTACAGTTACAACTGCTACACCCGTTGGGACCCCAGGTGTTCCAACTGTGGTAGGAGTACCCGGGGCAGTA
GGAACAGGTGTACCAACAGGAGCCGGAACAGCTGGTGTGGGTGCTCCTGGGGGTGGAGTGGGAGCACCAGTAGGTGCTCCGGGAGTAGTAACTCCGGGTGGACCTGTCACAGGAG
GCGGAGCTCCAACCGCAGGCGTACCAGGAGGTGCCGGTGTTGCTGGTGCGCCTGGTGTTGGAGTAGCTGCAGGAGCTCCCGGTGGAGGTACTACTACAGCCCCTGGTGCCGTACCG
ACCGCTGGGACGCCCGGAGGTGTGCCTGGAGTGCCAGGTGGTGGTGTAGGTGTAGTTCCCGTAGGTACACCAGGAGCTACGACCCCTGGAGGGACTCCTGGCGCTCCAGTCGGAGT
ACCTGGTGTCGGGGTACCTGCTGGAGGAACCACTGCAACAACAGTCACTCCTGGAACTCCAGGAACTGGTACTCCAGGTACAGCTCCTACAGGAGTAGGCACGGTAGGTGGCGTTC
CAGTTGGTGGGGCACCAGGGGGAGCGGGGGCTCCAGGAGCGACTGGCTGGCCG

[PGGVIL]40:
ATGAGCAAAGGACCAGGAGTACCCGGCGTTCCCATTGTAGTTGGAGTTCCTGGTTTAGTTCTTGGAGTGATTCCTGGAGGTTTGGGAGGATTGATACCTTTAATCATTGGAGTCTTAG
GAGTCCCTGTTGGCTTATTACCTCTAGGAGGTGGCCTTCCTGGTGGTCTAGTTATTGGACTCGGAGTTATAATAATAGGGGGTGGATTACCATTAGGTATACCCATAGGAGTTGGTATT
CTCCCTGGAATACTTGGTGTTCCAGGCGGTGGTATAATTGGTGGAGGAATATTAGGGATTCCAGGAGGCGTCCCAGGGGTAGGAGTGCCTGGGCTATTAGTTTTACCTGGCGGAGTA
ATTCCTGGCATTCCTGGGGGCGTAGTCATTCCAGGTCTGATAATTCCGTTAGGTGGTGGGGTTCCTGTCGGGGGAGGTATTATTCTACCAGGAATAGGCATCCCTGGTATCTTACCTAT
CGGTCTTCCAGGTGTCCCCGGTGTTGTCATACCAATAGGAGGACTACCTGGGATACCGGGCATACCTGGTGTGCTTCCAGTTGTGGGAGTAGGGTTACCAGTAGGTGTACCTGGAGT

ATTACTAGGTTTACCAGGTGGAATTTTGTTAGGACTTCCGGGTGTTATTCCTATAATTTTAGTAGGAATTCTTTTAGGATTATTGCCGATTGGAATTGGAGGGGTCGTACCAGGCTTGG GTGTAGTACCTGTAGGTATTCCCGGAGGTGGCTGGCCG
[PGGVIS]40: ATGAGCAAAGGACCAGGAATTATTAGCGTACCCGGCTCTGGTGTACCAGGCGGTGTTCCAGTAATCTCGGGTGGATCAGTACCTTCTGTCGTTGGTATAATTTCAGGAGGTTCACCA GGTGGAGGATCGAGTCCAGGTTCGGGAGGTGGTAGTCCGTCAGTTATAGTAGTTTCTGTTCCTATCGGAGTACCGGGGATTCCTGTTATAATCGGCATACCAGGATCTATAGGGGTA GGTATACCAGTTGGAAGTGGAGTCAGTGGGATTGGTTCCCCTGGCGGAGGGATAGGTTCTCCCATAGGATCACCTGGAATCATAGTACCAGGGTCACCAATTGTTGGAGGGGGCAG TCCTGTATCTGGTATTCCTGGAGGTAGTGGTATCTCAGGAAGCCCAGGTATTGGTATTTCATCAGGTAGCCCTGGGGGTGGGGTTGTAGTCTCTCCAGGCTCCATACCTGTGAGTGGT GTCCCAGGAGGCATCCCTGGAATTCCCGGTGGTGGAATAGGAATAATACCTATAGGAGTGCCTGGGTCTGTAGTTCCGGGCGGCGTTCCTGGCTCACCCATTGGGATACCAGGTATC GGTATACCCAGCGGGGGAGTTGTTAGTGTTGTCATTGTACCTGGAGTTCCAGGAGTAGGTGTGCCAGGTGTAAGTCCCGTAGGAATTGGAGTAATAGGTGGCATTCCAATAGGAGGA AGTCCTGGTGGTTCTGGATCTCCTGGTTCAGTCATAGGCTGGCCG
[PGGVKF]40: ATGAGCAAAGGACCAGGAGTAGGATTCGGAGTAAAACCAGGTGTAAAAGGATTTCCAGGTGGGGGAGTCTTTCCTGGTGTAGGTAAAAAAGTAGGTGGATTTCCGGTCGGCGGAG GTAAACCTAAAGGTGGTGGAAAGCCTGGTGGTAAACCAGTTAAAGGCTTCCCCGGCTTTTTTCCTAAAGGAAAGGGCAAACCATTTGGAGTTGGAGGCGTTGTTGGTGGCGGGGTC AAGAAGGGATTCAAGGGTAAACCCGGATTTAAATTTCCCGTGAAGAAAGGTAAGAAACCTGTAGGAGTGGGAGGATTTTTTCCAGGGTTTCCAGGCGGTGGTTTCGGTTTTGTAGTA CCAGGATTCTTTGTACCCGGTAAACCGTTTGGTAAGCCCGGTGTGCCTGGGGTAGGTTTCCCTGTCAAAGGTTTTGGTGTTGGAAAAGGTGTCGTTCCTGGCGTCAAAAAAGGCAAG CCAGGGGGCGTACCTGTTTTTAAAGGAGGAAAAGTCCCTAAGGGTGTATTTCCAGTATTTTTTGGGTTCCCAGGTAAATTTGGTTTTTTTCCGTTCGTAGGAGGGGTTGTAGTTAAACC TGGATTTGTTCCAAAAGGTAAAGGGTTCTTTCCATTCGGGGGTGGAGTACCTGGAGTGCCAGGTAAGTTTCCTGTGAAAGGAGTTCCAGGAGGTTTTCCTGGGTTTGGATTCCCTGGT AAAAAGTTCAAACCGGGATTTCCTGGAGGTGTTCCTGGCTGGCCG
[PGGVKY]40: ATGAGCAAAGGACCAGGAGTTGTATACCCAAAAAAAGGAAAACCGGGAGTACCTGGAGTAGGCTATTATCCTAAGGGCAAGGGAAAAGGTAAGCCTGGTTACAAATATTATTATA AACCTGGTAAGTATGGTTATCCATACGGAGGTTACTATCCTTATGGAGTGCCAGGTGTAGGTAAAGTCCCTGGCGGCGGAGTATATAAACCCTATGGGAAACCTTACGGTTATCCGG GTTATTATGTACCAGGAAAATATCCAGGCGGGGGCTATCCAGGATATCCGAAAGGAGTCTATCCTGGTGTTAAAGGATACCCTGGTGGTGTACCTGTAGGAGGGGGTGTCGTAGGT GGAGGGAAGCCGGGCGGTGGAGTTCCTGGAGGAAAGGGTGGAAAAGTTCCAAAAGTAGTAGGAGTAGTTGGAGTCGGTAAGAAAGTTGGTGGATATCCCGTTGGAGGTGTTCCTG TGTATAAAGGTGGCTACCCCGGTGTTTATCCCTATTATAAGAAGGGGAAAGTGGTAGGTTATCCTGTCTATGGAAAGCCAGGAGGTGGTTATGGTAAAAAACCTAAAGGAGGCGTA GTCAAGAAACCAGTAGGTGTGGGAGGATATTATCCAGGTTACCCAGGATACAAGTATCCAGTGAAGAAAGGTTATAAAGGCAAACCAGGTGGGGTTCCAGGGGTCAAACCTGTTG GATATGGTGTGCCTGGGGGAGGTAAACCAAAAGGTGTCCCAGGTGGCTGGCCG
[PGGVFS]40: ATGAGCAAAGGACCAGGATCATCAGGCAGTAGTGGCTCAGGTTTTTTCAGTGGTGGCGTTCCTAGTGGAGGGGGCGTAGTGTTTGTTCCATCATTTTTTGGTTTCTTTCCAAGTGGATC TGGTGGTGTAGTACCAGGCGTGCCCGGCGGGGGTGTAGGTGTTCCTGGTTTTGTTGGAGTTCCAGTAGGTGGTTTTCCTGTGGGAGTACCTGGAGTATTTGTACCATCTGTCGTAGGA GTCCCTGGATTTTTTGTGTTTCCAGGTGTACCGGGTGGTAGTTCTAGTCCTGGATCACCCTTCGGGGTCTTTGGCGGAGGATCGTCTGGAGGCGGTTTCCCTTTCGGCTCTTCACCTGGT GTTGTAGTCAGTCCAGGATTTGGATTCGGTGTCCCATTTGGGGTACCTGTAGGAGGAAGTGTCTTCTTTTTTCCGTTCGGAGGTGGGTTCCCCGGTGGATTTCCTTCTGGTTCATTTGGT GTGCCTGGGGGAGGTTTCGTTGTACCTTTTGGAGTCTCACCAGGGTCCCCTGGCTTCGTACCTAGTTTTGGTAGCCCAGGAGGGTTTGGAGGATTCTCTCCTTTTAGTTCAGGAGTTGG TTCGCCAGGTGGAGTGCCAGGAGTAGGGGTTGTATCTCCAGGCTTTCCTGGAGGTAGTCCATTCGGATTCCCAGGTTCTCCGGGAGGGAGCTTTGGTGGAAGTCCGAGTGTATTTGGA AGCGGCAGCCCTGGGTCATTTCCAGGCTGGCCG
[PGGVTY]40: ATGAGCAAAGGACCAGGACCAGTTACAGGAGGCACTCCAGGAGGTGTCGTCCCAGGCTATTATGTCTATCCTGGGGTCCCTGGTGGTACTCCTGGCTACGTACCAGGAACGCCTGG TACCTATCCTACGGGAGGGGTTCCAACAGGTGGTGGAACTGTACCTGGTACTTATGGTGTCCCCGGCGGGGGCTATGGAGTATACGGTTATCCAGGTACTGTTCCTGTTTATTACGGA TATCCCGGAGGTGGCTACCCAGGTGGAGGGACTCCCGGTGGTTATGGCGGCTATCCAGGCACCACTACTCCAGGTTACCCTGGAGGCGGTGTAGGAGGATATCCAGGATATGTACC AACTGGAGTTTATGTTCCCTATGGTTACGGTGGATACACTCCTTACACGACCGGAACTACTGGTACAGGTGTGGTCACTTATTATGGGGGAGGTTACTATCCAACGGGTACTGGAGGT ACACCTACAGGGGGTGGAGTCGTTCCAGGTGTTCCAGTAGGTTATTATCCTGGTGTAGTGACTCCTGGTTATCCTACTACTACATATACTACAACTCCTGTGTACGGAGGGTATGGAG GATACGTTGTGACAGGTTATCCGGGTGTTACATATGGATATGGTGGGTATCCAGTGGGATACTATGTTGTACCGGGCGGAGGAACTCCGACTGGTGTTGGTGTGCCAGGTACCCCTG GCGTACCTGTGGTAGTTACTGTAGTTGTTCCTACCGGCTGGCCG
[PGGAKY]40: ATGAGCAAAGGACCAGGAGGCAAGGCAAAAGCATATCCTGGATACTACCCAGGATACGGAGGGGGTTATCCCTACGGGGCCTATGGAGGCTATCCAGGTGGTGGTTATTACAAAC CAGCCGGTTATCCGTACAAAGCTAAGAAGTACAAGCCCGCTGGAAAGCCTGGTTACCCAGCTAAAGGAGGTGGAAAAAAGAAATATCCAGGGAAGGGTAAACCTAAGTATGGAG CGCCAGGGGGAGGGAAATATGGTAAAGCAAAAGGTGCTAAAGCTTACGCAGCAGCCTATCCTGGCTATGCTCCTAAATATGGAAAACCTGGGAAAGGAAAACCAGGAAAGTATC CTAAGGGAGCAGGTAAGCCTTACGGTGGAGGATACCCGGGAGGTTATGGGAAACCTGGAGGTGCGCCGGGCGCGGGAGCTCCTTATGGTGCATATGGTTATCCAGGCGGAGGAAA AGCAGGCTACCCTGGGGGTGGAGCTGGATACCCCGCAGGTTATCCTGGTGGGGCGGGTGCTTATTATGGATATCCCGGAGGTAAACCGAAAGCGTATGGTGGATATAAACCTTATA AGAAAGGTGGGAAGCCAGGTTACAAAAAACCGGGTGGTAAACCAGGTGCAGCGAAACCCGGATATCCTGCCGGAGCTCCAGGAGCCCCCGGTGGCGGTGCTGGAGCCGCTCCTG CTGGCAAACCTGGTGCTCCAAAGGCAGCTGGTGCCCCTGGTTATGCAGGGGGCTGGCCG
[PGGAFS]40: ATGAGCAAAGGACCAGGAGCATTCGCAGCAGCATCTCCAGCAGGAGCTCCCGGTGCTAGTCCAGCGGGAGGATTTCCAGGGTTCTTTGCACCAGGCAGCCCAGGTGGTGCGCCTGC TTTTTCAGGAGGATCGGCTCCTTCATCGGGTGGTGGTTTTGGGTCACCCTTTGGTGCTTCACCTGGAGCGTCGGGATTCCCTGGTGGCGGAGCTTCGAGTGGTTCTCCTGGGGCGCCAG GTTCGTTTCCTGCATCAGGAGCGCCGGGCGGTGGATTCGGAGCACCTGGTAGTGGAGGCGGGTCCCCTGCAAGTGGATTTCCTGGTTTTTTTGGAGGTGCCTTTCCAGGCTTCCCAGC TTTCTTCGGTTTTCCTGGCAGTTTTGGTAGCTCTCCGTCAGGTGGAGGAAGTCCAGGAGGTGGGTTTGGTGGAGCAGGTTTTTTCCCATCTGGAAGTGGCTCAGGATTCAGTGGGAGTC CTGGATTCTCATTTCCAGCTAGTTCTGGTTCAAGTCCTGCGGGGGCCGGAGGTTTCTTTCCCGGAGCCTCAGCTCCAGCCGGTGGGGGAGCGGCAGGAGGGGGTGCTGGTGCCCCGG GTGCATTTTTTGCTCCCTCTGGATCACCAGGATCTGCAGCTGGTTTCGCTGCACCTTTTGGTTTTCCAGGTAGTAGTTTTAGCCCTGGCTTTCCTGGAGGGTTTCCAGGATTTGGATTTTT TCCATTTGGAGCCCCTGGGTTCCCAGGCGGCTGGCCG
[PGGAFT]40: ATGAGCAAAGGACCAGGAGCAGCAGGAGGAACAGGTTTCTTCACAGGAGGGGCGCCAGGAGGTACTCCAACAGCATTTGGAGGTTTCACTCCGTTTACTTTCGCTGGAGGCGGTGC TCCAGGCTTCTTTGCGTTTCCAGGAGCGCCGGGTACCTTCCCAACTTTTGGTGCCCCCGGTGCAGCTCCTTTTGGAACGACCCCAGGCACTTTTTTTGGGTTCCCTGGAGGGACACCTG GCACACCAGGGGCTGCGACTCCAGGTTTTTTCGGAGCCTTCGGATTTCCTGGTGCTTTTGCAACATTTGGCTTTCCAGGGGGTGGAGCTGGTGCGGCTCCCGCAGGTACGCCCGGAGC TCCCACGGCAGCTGGAGCCACAACCCCTGGGGCCGGTGGTACAACTGGTGGTGCACCTACGGGAGGTGGATTTGCTCCTACATTCTTTGGATTTTTTCCTACTGGAACCGGAGGTGC CGCTCCAGGTGCGCCTGGTGGGGGTACTACTACCTTTCCTGGAGCTACTCCCTTTGGTTTTCCAGGTACAGGAACTCCTGGTACGTTTGGAGCTCCTGGGGGAGGGTTCTTTCCGGGA GCAGGGTTTGGTGGCGGAGCACCCGGGGCTCCTGCAGCAGCGGGTGGTGGAACACCTGGTACTGCTACACCAGGTACTTTTCCGACAGGTGCTGGGACTTTTCCCGGTACATTTCCA TTCGGTGGAGGATTCCCAGGTGGTTTCCCTACCGGCTGGCCG
[PGGILS]40: ATGAGCAAAGGACCAGGAAGCATAATAGGATTACTTCCTCTTGGTGGATCACCTGGAGGCATCCCCGGAATTTCATTACCTGGATCTGGATCGCCTGGGTCACTTCCAAGTGGATTG CCTGGAATTCCCGGTGGAATCAGGTCTCCTGGTTTATCTAGTCCCTTAGGAGGTGGTATACCTATTGGAGGTATCAGCCCTGGCATCATTCACCGTCGGGAG GTCTACCTGGCTCACCAGGCTCTCCAGGAATTCTTCCTATCATAGGAATAGGTTTACCTATCGGTATTCCAGGTCTGTTAATACTGCCGGGCATTCCTTCTATAATTGGTATCCCTGGA TTAATTCTCGGAATAAGTCCTGGTGGTTTGGGAGGGTTGAGTCCGTTACTAGGATTGTTACCTTCGGGTGGTGGCCTTCCATTAGGTGGAGGACTACCAGGTGGGTTAATTAGTGGTCT AGGAATCTCTTCAAGCGGCGGGGGTAGTTTATTAGGCTTACCGGGTATATCTCCATCATCTTTGATTGGAATATTATTACCATCTGGTTCAGGAGGAATAATACCAGGATTAGGAATT ATTCCCATTGGTTCCCCAGGAATCCCTATATCAATACTTTTAGGTCTTCCAGGAGGGAGTCTAGGTAGCCCAGGGGGAGGATCAGGTATAGGATCTTTACCAGGGTCGTTAGGTATTC CTGGCGGAGGTTCGAGTGGGGGCGGTTCTGGCTGGCCG
[PGGILT]40: ATGAGCAAAGGACCAGGAACTATTTTACCAGGCATTGGACTTCCTGGAGGATTACCATTAACAATTGGTGGAATATTACCAATACTTCTAGGTACTACAATAGGTGGCATTCCAGGT GGAGGGATTATTCCTGGTATAGGGGGCGGTATCCCCATATTAACATTATTAATCTTACCTACTGGCTTACCTGGGACTCCATTGACAACAGGAACCCCAGGAATCACTGGATTAGGT ACTCCGTTAGGAATAGGATTGATTGGAACAGGTATTCCAGGCGGGGGAACTCTGGTATTCCCACTGGAATCCCTGGTACAACTTTGCCAGGAATTCCAACTTTGGGTGGTGGTATACCC TTAATAGGCACTCCAGGAGGAATTGGTACGATAGGCATCCCAGGGACTGGAACAATTATAGGTATACCGGGTGGAGGTTTGTTACTCATTCCTGGAATCACACCGTTGATAGGACTT CCAGGGGGTACTCCAGGTACACCTGGTGGAGGCACAGGAACGACTCCTACAGGGCTCCCTGGAATTCTGCTGCCCGGTGGTCTTCCCGGCATACCAACTGGAACTCCTGGTACTGG GACACCCATAGGAGGTTTACTTATATTATTAACTCTCCCTGGCCTTCCAGGTCTTGGATTACCTGGTTTAATACCTTTAGGAACAGGCTTGACGGGAGGTACGCCGACAGGAGGCATA CCCGGAGGTATCGGTATTCCGGGAATTTTAACTACTGGCTGGCCG
[PGGIKY]40: ATGAGCAAAGGACCAGGAAAATATAAATACCCCAAATATGGGATTCCAGGAGGTGGGTACGGAATTCCCTATGGAATACCGGGTTATCCGAAAATTATCTATGGTATAGGTGGTGG TTATATAATTGGAATTCCTGGAGGCAAAAAATATCCAAAGATAGGAAAACCGGGAGGTTATCCTGGATACGGGTATCCAATCGGGGGCGGCATCCCAGGTGGAATAGGTAAATAC ATTGGTGGAGGTAAAAAGTATGGAGGATACTATCCTTACGGTAAGCCAGGTATAAAAAAACCTGGTGGCAAGCCGGGGATTTATCCTAAGATCGGTTACATAGGAATACCAGGTAA AGGGAAGCCTGGAAAAATTCCTAAAGGTTATGGTAAACCCTACGGCGGAGGAAAGAAAAAGATTCCTGGCATTCCTTATGGTTATCCAGGTTACTACAAACCAGGCATACCAATAG GATATATAGGAGGAATAAAACCAATCAAGAAAGGTAAGATACCAAAAATAGGCTATCCAGGGGGTGGTATACCTTATAAAGGAGGCGGGAAATATCCTGGTTACCCTGGCGGTGG

ATATGGTGGGAAACCTAAATATATCGGAGGTATCCCTGGGGGAGGGATCATCATTCCAGGTATTGGAGGGGGAATACCTATCAAATATAAGAAGATTAAACCTTATGGAAAGCCCG GATATCCCAAGTATTATGGCTACCCAGGAATATATGGATATCCTGGCTGGCCG
[PGGLKF]40: ATGAGCAAAGGACCAGGACTCGGAGGATTATTAGGTGGCGGTCTTAAAAAAGGCAAGCCAGGGGGATTACCATTATTTAAAGGAGGAAAATTACCTAAGGGTCTATTTCCTCTATT CTTTGGATTCCCCGGTAAGTTTGGTTTTCTACTTCCAGGATTTTTTCTGCCCGGCAAACCATTTGGTAAACCAGGGCTACCTGGGCTCGGTGGTCTTCCATTCGGGGGTGGACTTCCTG GCCTACCAGGTAAATTTCCACTTAAGGGACTGCCTGGGGGCTTTCCTGGTAAAAAATTTAAACCCGGATTTCCTGGAGGACTACCTGGACTAGGTAAAAAGTTAGGAGGCTTCCCTC TTGGTGGAGGGTTTTTCCCTGGGTTCCCAGGAGGTGGATTCGGCTTTTTTCCATTTTTAGGATTTGGACTCAAACCTGGTCTGAAAGGATTTCCAGGCGGAGGTCTTTTATTAAAACCG GGATTTCTACCCAAAGGTAAAGGATTCTTCCCGTTTGGAGGTAAACCTAAAGGTGGTGGTAAGCCTGGTGGGAAACCTCTCAAGGGGTTTCCAGGGTTCTTTCCAAAAGGAAAGGGT AAACCGTTCGGACTTGGATTTCCGTTAAAAGGTTTTGGTCTAGGAAAAGGTCTGTTGCCAGGTCTAAAGAAGGGATTCAAAGGAAAACCTGGATTTAAATTCCCATTGAAGAAAGGT AAGAAACCACTAGGATTGGGAGGTTTTTTTCCTGGCTGGCCG
[PGGSTY]40: ATGAGCAAAGGACCAGGAAGCACTTACAGTTACGGCTCTCCTTCTGGATACTATCCATACGGAGGGGGCTACCCAGGTGGATATCCGGGCGGAGGATCAGGTTATCCAAGTGGTTC ACCTGGTAGTTATTATGGTTATCCTGGGACCCCTGGCACGTATCCTACAGGTAGCGGCACACCTGGAACGCCAGGTAGTCCAGGAGGGTCAGGTACGTCTGGTACATCGAGTGGTGG CTCAAGCCCAGGCAGTGGATCACCAGGTTACTATTCATATCCTGGAAGTCCTGGAGGCGGTTCAGGATCTAGTCCTAGTGGAACTCCTGGCTATCCTTATGGGGGAGGTACAACAGG AGGTGGTTACCCTTACGGTACACCGACTGGTTCGGGAACTTATCCAGGTACTTATGGCAGCCCTGGTGGAGGAACTACAAGTTCAACTTCGACACCAGGATCGCCTGGTACCTACCC TGGATCTTCTACTCCAGGCTATGGTAGTGGGTCTTACCCCGGCGGGGGTTATCCGAGCTATACTACTACATATTCTCCAGGTTCTCCTGGTTATGGAACCTATCCAGGAACACCCGGT TATTATCCCTATGGATATGGTGGGAGCACCCCAGGGACAGGATACCCTACGACGGGAGGATATCCCGGTGGTACTCCTACTTCATATGGAGGTTATACACCTTATACAACTGGAACA ACGGGTACTGGATATTACACTGGAGGGAGTCCAACCGGCTGGCCG
[PGGXXX]40, X=VALK: ATGAGCAAAGGACCAGGACCAGGCGTAAAACCCGGCGGTGGAGCATTAAAGAAGAAACCGGGAGCTCCTTTAAAAGGTTTGAAACCAAAAGGAGGTGGTGCAAAAAAACCTGTC GGAGCTCCAAAGAAAGGATTGAAGGGAGGTGCCGTCGTGGTAGGTGCAGGAGCCAAAGCGAAACCTGGGTTGCCTGGAGCCGGAGGTGTCTTAGGATTACTTCCAGGAAAGCCAG GAGGACTTCCTAAAGTAGTAGGTCTGAAAGGTGGCGGAGCCCCCGGTGGTGGTCTAAAAGCTGTTCCAGTAGGTGTTCCTGTAGGTAAACCTAAACTTGGTAAAGCACCTGGAGTTC CTGGTGCGCCCGGAGCAAAACCAGGAGTCCCTGGAGGCGCGGCCCCTGGTAAGCCCGTTGGAGTACCTTTGGGTGGAGGTCTAGCTCCATTAGGACTAGGAGCACCATTGGGAGCA GCTAAACCAGTTGTTGGTGCCCCAGGAGCTGGAGGAAAACCACTCGGTGCTGCAGTACCAGGTGTGGGTCTTGGAGGGCTACCTCTAGGTGGGGGTGTACCTAAGTTAGGAGGGGT TGGGGTAGGAGCGCCAGGGTTACCTGGTGGACTACCTGCTGGAAAAGTAGGAGTAGGGGGAGGATTAGTTGGATTAGGGAAGCCTGGGAAATTACCCGCTGGTGTAGCACCCGTA GGCGGGGCTGCCTTGCCAGGTGGAGTCGGTAAATTAGTATTACCTGGCTGGCCG
[PGGXXX]40, X=VALKT: ATGAGCAAAGGACCAGGACCAGGAAAAACACCAACTCTCGTAGGAGCGGCCCCCGGTGGAGGCGTCGGTCTGACTCCAGGTGGCGCAGCAGGAGTCGTTCCTAAACTTGGATTGG GAGGTACGCCGGGCGGGGTAACTGTTGTTCCAACTGCAACAGGTGGTGGAAAGAAACCTGGTAAACCGAAAGGTACAGCTCCTGGCACTGGAGGGGGAGTTCCGGCAGCCGTATT GAAACCAAAAGGAGGCAAGCCTGGCGGTGTGCCTGGAACTACTCCTGGTGGGACACTTCCTGGAGGTCTACCTGGAAAACCCGCAGGTAAAAAACCTAAAGGTTTAGTTCTTCCTA CTGGACTCGCTGCTGGGGTTCCTGGGGGTGTATTACCAGCGGGAGGACTAGCTCCAGGGCTTCCAGGAGGGAAACCAGCCAAGGGAGGGTTACCTCTTGGAGGATTGTTAGGTGTT GGAGTAAAACCGGGTGGTAAGACACCTGGGAAAGGAGCACCTGTCGGAGTAACTCCTGTGACGGGAACAGTACTACCAGGGACCCCTGGGTTGAAGGGTGCCCCTGCAGCTGGTG GTCTAACGGTAGGTAAATTAGCTGGAGTTGCTACATTGACAGTCGCACCTACAGGAGGAAAAGGTGGAGTGGGTACACCAGGTAAGCCAGGTCTTAAACCAGGCGGAGCAGTACC AGGATTAGGAGGTGGGGCTCCCGGAGGAACTGGTGCAGGCAAACCTGGCTGGCCG
[PGGXXX]40, X=VALKTI: ATGAGCAAAGGACCAGGAGGACTACCAATCGTTAAAAAGAAACTTCCTGGAACCCCTGGTGTAGGAGGTGGGATCATAGTAGGGACATTAGTCAAACCTAAGGGAGGGGTAGCA CCAGGTCTTGCAGGAGGCACTTTACCGGGTGGTCTACCTGGAATACCAAAAGGTAAGAAGACTGTTTTACCAGGATTGGGAGGAAAACCAATAGGAATTGTTAAACCGGGCGGAG GAACACCTGGAAAAGTTCTACCAGTAGGTGGTGGAGCGATAACTGTAGCTCCAGGAGTAGGCAAACCAGGTACGGGAGCACCCATTCTAGGAGTTCCTGGGGCCACTGGAGGCGG TGCATTACTTCCAGGGGGTGTGATACCTGCCGCTGGTGGCATCCCTGGCGTAGGTACTCCCGGTGGAAAGATTCCGGGTATAAAACCTGGAGGATTAGGATTACCTGGCACAGGAG GGAAGATCGGAACGCCGGGGGCTGGAGGTGCGACTCCAGGCGCATTGAAAACTATAGGAGCCGGAGGTAAACCTGCGGGAGGTACAACACCGACCGGAGCTTTAATAATTACAG GGGCACCTGGGTTGATTCCCGGAGTAGTTGGTGGGGGAGTGACCCCAGGGTTACCTGTTGGAGCAGCTGGGACTCCTTTAATTGGAGTATTACCAGCAGGAAAGCCAGGAGCTGTA CCAGGCATACCAGGTGGAACTATTCCTGGTGGTAAGCCTGGGAAACCTGGCTGGCCG
[PGGXXX]40, X=VALKTIY: ATGAGCAAAGGACCAGGACCAGGCACTTACCCCGGCTACGGATATGTTTATCCAACGACTGGAGGCATTCCTGGGGGTGTTGTTCCGGGCGGAGGAACTAAATTAGTACCAGGTAA AGGTAAGGGAGGTGCTAAAGCTCCGGGTACTGTTCCAGTAGGAGCTGGTGGAGGAAAGATCGTACCTATCTATGGAATTGCTCCTGGTAAATACGGTTACCCAGGTGGTGGGATTG TCCCTGGTATTACAACTCCAGGGCTGCCAACAGGAAAAAACCTTATGGAGGGGTTCCCGTTCTATATGGTAAACTTCCAGGCGCACCCGGTATTCCTACAGCGGGGGCACCAGGTT ATATCGCTCCAGGAGTACCTGGAGGTCTAGTGAAAGGTGGCACAGGTATAGCTCCATTAGGAGTCATTTTAGTATATATTGGAGTAGGTGGAATCAAGGGTGGAGCATTACCAACTG GTGGTTTATATCCAGGTGCAGGTACACCTGGTTATCCTGTCGGAGGTGGGGCTCCAGCAGGAGGTATTGCATTGAAACCTGGAATAACACCAGGAACTGCGGCCCCTGGGCTTCCTG GAAAAGGAGGGAAATATACATACCCTGGCGGTGCTCCCGGTGGTACAGGAGGCGTCCCAGGCTATCCTGGACTAGCGTTAAAGTTAGGTATCCCTACAAAAGGGGGAGGAATCGG TCTTCCTTATATAGGATTACTACCTGGCAAGCCAGGAGGGGGCTGGCCG
[PGGXXX]40, X=VALKTIYS: ATGAGCAAAGGACCAGGATTACCAGGCGGAAAAGGTGTAGGGTCAGGAAGCCCAGCCGGTGTTACTACTCCAACAGGAATACCTGCTGGCCTAGCTCTACCTGGCTACTATCCAG GTGTGGCAGGGTTGCCAGGGACGCCGGGCGGTGGTACGATAGGAAAGTCAACTGTTCCAGGTTACACTCCTAGTGGTGGTTATAGTCCAGGAGGACTCGCTTACCCTGGAGGAACG GGTAAAGGAGGGGTACTTTTACCAATCTATGGTGGAGCAGTTCCTGGCGCTGGTGGGAAACCTATTTTGGGTGGAATCAAAACACCAGGAAAAGCTCCAGCTGTAGGAGGTCTGAT CCCAGGGGGTGTCATTCCTGGAAAGATACCAGGTCTACCATTGGGAGGTTCTCCAGGTACAACTCCTGGGGTTCCATTATATGGAGGTGGTCTTGGTTTAGGTGGGGGCATTCCAGG CACTGGATCTCCTTATACAGGTGGATTAGTTAAAGGTTATCCATCTGGAAAACCCGCAGGCTATCCTAAAATTGGTGGCGGCAAGTATGTACCGGGAATTGGAGGCATCATACCTTA TGGAAGTCCGAAGGGTGGTGCAAAAGCACCCGGAGGTATCAGTCCTGGAGTCGGGGGAGGATACCCAGGTAAACCGTATGCTGCAGGAGCGGGAGTATATTCAATAGTACCTACT GGTACTCCGTCTGTAAGTGGAGGGAAGAGCCCTGGTTCACCTGGCTGGCCG
[PGGXXX]40, X=VALKTIYSF: ATGAGCAAAGGACCAGGAACCCCATTCGGATACCCCGGTGGAACGCCAGGGTCAGGAGTTGGTCTGAGCCCAGGTTATGGAGTAACAGGTCTAGCGCCTGGGAAGAGTCCTGGCT TTGGTTTTGGTAAAGGTTCGACTCCAGTAGGATTAGGAAGTGGTTTCGGTGTACCTATAGGATATTTTCCCGGTAAATATTATCCTCTAGGTAAGACACCAGGAGGATCTCCAGCCTA TGTAGGTGGAGGACTTGGTGGTGGATACGGAACTCCAATTATCGGAGCCGGAGGAACTGGTGGCTATCCATCAGTGATTGGATTTGGAGGTAAACCGTCTGGTAAGCCTGGTACTTT TCCTGTTGGAGCGTTTCCGGGTACTCCTAAGGGTAAACTACCTGGCAAACCTATTGGAAAGATAGGAACACCCGCAGGGGTCGTACCATTGAAAGGTGGGTTAGTACCTGGTTACGC ACCAGCGGGAGGTGGTTCTGCTCTGCCGGGCGGTGCCCCTAAAGGAGGCGGAGTGGGAGCTGGCCTTCCATTAGGTGTGACAATTCCAGGCATTCCTGGATTTCCAGGGGGATATA GTCCAGGTGCAGGAGCATACCCAGGATTTAGCCCTGGGGGTGGTACAATAATACCTGGAGGGTTTGGATACATCCCAGCAGGTTTTCCACTAGGAGGTTACATATTCCCTGGTATAT TGGGTGCTGTTCCAGGATCATCTGTCCCAAAAACATTAGGCTGGCCG

Table 16: Summary of published IDPP phase transition data from various sources

Protein name	Tt (C)	Length (amino acids)	Concentration (uM)	Molecular Weight (Da)	Source (DOI/Patent)
[VPAVG]45	41.9	237	50	13445	10.1038/nmat4418
[AVPGVG]25	66.8	162	50	34620	10.1038/nmat4418
[TVPGVG]65	51.8	402	50	15848	10.1038/nmat4418
[VPGAVG]30	56.6	192	50	25459	10.1038/nmat4418
[GVPGVA]50	45	312	50	16254	10.1038/nmat4418
[VPGVA]35	42.9	187	50	13715	10.1038/nmat4418
[VPGVG]30	49.8	162	50	15097	10.1038/nmat4418
[VHPGVG]25	36.8	162	50	13715	10.1038/nmat4418
[VGPVG]30	51.7	162	50	13445	10.1038/nmat4418
[VPAGVG]25	70.5	162	50	21855	10.1038/nmat4418
[VPTGVG]40	54.1	252	50	18250	10.1038/nmat4418
[APVGVG]35	59	222	50	21214	10.1038/nmat4418
[APVGLG]40	31.9	252	50	22415	10.1038/nmat4418
[TPVAVG]40	38.9	252	50	21014	10.1038/nmat4418
[VPHVG]40	29.3	212	50	11042	10.1038/nmat4418
[VGPAVG]20	71	132	50	17169	10.1038/nmat4418
[TPVAVG]30	44	192	50	16254	10.1038/nmat4418
[VAPVG]35	47.3	187	50	9901	10.1038/nmat4418
[VAPVG]20	56	112	50	14546	10.1038/nmat4418
[VTPAVG]25	73.9	162	50	14136	10.1038/nmat4418
[VAPVG]30	44.1	162	50	18371	10.1038/nmat4418
[VAPVG]40	40	212	50	14196	10.1038/nmat4418
[TVPGVG]25	76.1	162	50	20489	10.1038/nmat4418
[GVAPV]45	38.2	237	50	13715	US20120121709A1
[GVGPV]30	51.2	162	50	13445	US20120121709A1
[GVGAPV]25	68	162	50	14196	10.1038/nmat4418
[VPTGVG]25	77.3	162	50	12363	10.1038/nmat4418
[VHPGVG]20	47.9	132	50	11923	10.1038/nmat4418
[TPVAVG]20	50.1	132	50	71402	10.1038/nmat4418
[[GAVPGV]5[GCVPGV]]24	31.8	876	50	71402	US20120121709A1
[[GVPAGV]5[GVPCGV]]24	32.5	876	50	105527	US20120121709A1
[[VPGVG]0.86[VPGWG]0.14]250	8.458904	1250	400	104569	10.1002/bip.360320913
[[VPGVG]0.9[VPGWG]0.1]250	11.28425	1250	400	105658	10.1002/bip.360320913
[[VPGVG]0.8[VPGYG]0.2]250	8.0879	1250	400	104633	10.1002/bip.360320913
[[VPGVG]0.86[VPGYG]0.14]250	12.7968	1250	400	103865	10.1002/bip.360320913
[[VPGVG]0.91[VPGYG]0.09]250	16.621	1250	400	106427	10.1002/bip.360320913
[[VPGVG]0.67[VPGFG]0.33]250	6.14726	1250	400	104842	10.1002/bip.360320913
[[VPGVG]0.8[VPGFG]0.2]250	12.62557	1250	400	103593	10.1002/bip.360320913
[[VPGVG]0.9[VPGFG]0.1]250	18.33333	1250	400	103342	10.1002/bip.360320913
[[VPGVG]0.9[VPGHG]0.1]250	20.55936	1250	400	102392	10.1002/bip.360320913
[VPGVG]250	23.98402	1250	400	102953	10.1002/bip.360320913
[[VPGVG]0.84[VPGLG]0.16]250	20.98744	1250	400	103485	10.1002/bip.360320913
[[VPGVG]0.69[VPGLG]0.31]250	18.27626	1250	400	103654	10.1002/bip.360320913
[[VPGVG]0.64[VPGIG]0.36]250	19.1895	1250	400	103205	10.1002/bip.360320913
[[VPGVG]0.77[VPGIG]0.23]250	20.78767	1250	400	103891	10.1002/bip.360320913
[[VPGVG]0.8[VPGEG]0.2]250	69.24658	1250	400	103157	10.1002/bip.360320913
[[VPGVG]0.81[VPGDG]0.19]250	52.75114	1250	400	102679	10.1002/bip.360320913
[[VPGVG]0.93[VPGDG]0.07]250	35.05708	1250	400	103844	10.1002/bip.360320913
[[VPGVG]0.8[VPGKG]0.2]250	42.87671	1250	400	103031	10.1002/bip.360320913
[[VPGVG]0.91[VPGKG]0.09]250	31.03311	1250	400	105408	10.1002/bip.360320913
[[VPGVG]0.59[VPGQG]0.41]250	39.02397	1250	400	103900	10.1002/bip.360320913
[[VPGVG]0.79[VPGQG]0.21]250	31.57534	1250	400	105872	10.1002/bip.360320913
[[VPGVG]0.76[VPGRG]0.24]250	32.51712	1250	400	105277	10.1002/bip.360320913
[[VPGVG]0.64[VPGMG]0.36]250	23.24201	1250	400	102612	10.1002/bip.360320913
[[VPGVG]0.78[VPGCG]0.22]250	25.2968	1250	400	98899	10.1002/bip.360320913
[[VPGVG]0.67[VPGGG]0.33]250	34.42922	1250	400	100288	10.1002/bip.360320913
[[VPGVG]0.8[VPGGG]0.2]250	29.86301	1250	400	101391	10.1002/bip.360320913
[[VPGVG]0.67[VPGSG]0.33]250	32.08904	1250	400	101765	10.1002/bip.360320913
[[VPGVG]0.79[VPGSG]0.21]250	29.52055	1250	400	100512	10.1002/bip.360320913
[[VPGVG]0.73[VPGAG]0.27]250	29.86301	1250	400	100989	10.1002/bip.360320913
[[VPGVG]0.8[VPGAG]0.2]250	28.29338	1250	400	101045	10.1002/bip.360320913
[[VPGVG]0.81[VPGAG]0.19]250	28.03653	1250	400	102453	10.1002/bip.360320913
[[VPGVG]0.88[VPGTG]0.12]250	27.38014	1250	400	103110	10.1002/bip.360320913
[[VPGVG]0.81[VPGNG]0.19]250	28.89269	1250	400	102661	10.1002/bip.360320913
[[VPGVG]0.93[VPGNG]0.07]250	25.86758	1250	400	15865	10.1002/bip.360320913
[VPGAG]40	105.2	206	25	31122	10.1021/bm4007166
[VPGAG]80	83.8	406	25	61637	10.1021/bm4007166
[VPGAG]160	65.4	806	25	16097	10.1021/bm4007166
[[VPGVG]0.1[VPGAG]0.9]40	99.50976	207	25	31467	10.1021/bm4007166
[[VPGVG]0.1[VPGAG]0.9]80	76.28833	407	25	62206	10.1021/bm4007166
[[VPGVG]0.1[VPGAG]0.9]160	57.445	807	25	16090	10.1021/bm4007166

[[VPGVG]0.2[VPGAG]0.8]40	89.42568	206	25	31571	10.1021/bm4007166
[[VPGVG]0.2[VPGAG]0.8]80	66.72	406	25	62535	10.1021/bm4007166
[[VPGVG]0.2[VPGAG]0.8]160	51.62	806	25	16322	10.1021/bm4007166
[[VPGVG]0.3[VPGAG]0.7]40	82.76643	207	25	31916	10.1021/bm4007166
[[VPGVG]0.3[VPGAG]0.7]80	58.92	407	25	63104	10.1021/bm4007166
[[VPGVG]0.3[VPGAG]0.7]160	46.77	807	25	16434	10.1021/bm4007166
[[VPGVG]0.4[VPGAG]0.6]40	77.65333	207	25	32141	10.1021/bm4007166
[[VPGVG]0.4[VPGAG]0.6]80	53.22	407	25	63553	10.1021/bm4007166
[[VPGVG]0.4[VPGAG]0.6]160	42.59	807	25	16546	10.1021/bm4007166
[[VPGVG]0.5[VPGAG]0.5]40	70.41167	207	25	32365	10.1021/bm4007166
[[VPGVG]0.5[VPGAG]0.5]80	48.695	407	25	64002	10.1021/bm4007166
[[VPGVG]0.5[VPGAG]0.5]160	39.245	807	25	16763	10.1021/bm4007166
[[VPGVG]0.8[VPGAG]0.2]40	54.43	206	25	32918	10.1021/bm4007166
[[VPGVG]0.8[VPGAG]0.2]80	38.225	406	25	65229	10.1021/bm4007166
[[VPGVG]0.8[VPGAG]0.2]160	30.845	806	25	16987	10.1021/bm4007166
[VPGVG]40	44.495	206	25	33367	10.1021/bm4007166
[VPGVG]80	32.3	406	25	66127	10.1021/bm4007166
[VPGVG]160	26.445	806	25	20567	10.1021/bm4007166
[VPGIG]48	22	242	25	76489	10.1016/j.jconrel.2013.05.013
[VPGSG]192	56.5	961	25	40896	10.1016/j.jconrel.2013.05.013
[VPGIG]96	11.5	482	25	39550	10.1002/pro.2063
[VPGVG]96	31.9	482	25	38392	10.1002/pro.2063
[VPGSG]96	62.6	482	25	19894	10.1002/pro.2063
[VPGVG]48	42.4	242	25	24203	10.1002/pro.2063
[[VPGVG]5[VPGAG]2[VPGGG]3]6	62.1	307	25	35941	10.1021/bm015630n
[[VPGVG]5[VPGAG]2[VPGGG]3]9	50.8	457	25	47679	10.1021/bm015630n
[[VPGVG]5[VPGAG]2[VPGGG]3]12	45	607	25	59417	10.1021/bm015630n
[[VPGVG]5[VPGAG]2[VPGGG]3]15	42	757	25	71154	10.1021/bm015630n
[[VPGVG]5[VPGAG]2[VPGGG]3]18	41	907	25	94630	10.1021/bm015630n
[[VPGVG]5[VPGAG]2[VPGGG]3]24	38.4	1207	25	129843	10.1021/bm015630n
[[VPGVG]5[VPGAG]2[VPGGG]3]33	36.2	1657	25	48991	10.1021/bm015630n
[[VPGVG][VPGAG]8[VPGGG]7]8	77.7	647	25	61057	10.1021/bm015630n
[[VPGVG][VPGAG]8[VPGGG]7]10	70.8	807	25	97254	10.1021/bm015630n
[[VPGVG][VPGAG]8[VPGGG]7]16	63	1287	25	121385	10.1021/bm015630n
[[VPGVG][VPGAG]8[VPGGG]7]20	60.5	1607	25	8918	10.1021/bm015630n
[VPGVG]20	76.4	107	25	13013	10.1021/bm015630n
[VPGVG]30	50.4	157	50	25298	10.1038/nmat4418
[VPGVG]60	34.8	307	25	37582	10.1021/bm015630n
[VPGVG]90	30.6	457	25	49867	10.1021/bm015630n
[VPGVG]120	28.6	607	25	23892	10.1021/bm015630n
[[VPGKG][VPGVG]6]8	59.6	287	25	47056	10.1021/bm025671z
[[VPGKG][VPGVG]6]16	44.8	567	25	93384	10.1021/bm025671z
[[VPGKG][VPGVG]6]32	39.2	1127	25	14709	10.1021/bm025671z
[[VPGKG][VPGVG]16]2	56.9	177	25	21699	10.1021/bm025671z
[[VPGKG][VPGVG]16]3	44.8	262	25	42671	10.1021/bm025671z
[[VPGKG][VPGVG]16]6	33.4	517	25	84613	10.1021/bm025671z
[[VPGKG][VPGVG]16]12	29.3	1027	25	18101	10.1021/bm025671z
[[VPGQG][VPGVG]6]8	47	217	25	35473	10.1089/ten.2005.11.1768
[[VPGQG][VPGVG]6]16	36	427	25	52846	10.1089/ten.2005.11.1768
[[VPGQG][VPGVG]6]32	32	637	25	14709	10.1089/ten.2005.11.1768
[[VPGQG][VPGVG]16]2	56	177	25	21699	10.1089/ten.2005.11.1768
[[VPGQG][VPGVG]16]3	45	262	25	42670	10.1089/ten.2005.11.1768
[[VPGQG][VPGVG]16]6	31	517	25	84613	10.1089/ten.2005.11.1768
[[VPGQG][VPGVG]16]12	30	1027	25	15778	10.1089/ten.2005.11.1768
[[VPGKG][VPGVG]7[VPGFG]]4	53.9	187	25	30828	10.1021/bm060849t
[[VPGKG][VPGVG]7[VPGFG]]8	36.8	367	25	60928	10.1021/bm060849t
[[VPGKG][VPGVG]7[VPGFG]]16	30	727	25	14448	10.1021/bm060849t
[[VPGKG][VPGVG]2[VPGFG]]8	66	167	25	28169	10.1021/bm060849t
[[VPGKG][VPGVG]2[VPGFG]]16	60	327	25	55610	10.1021/bm060849t
[[VPGKG][VPGVG]2[VPGFG]]32	44.4	647	25	18324	10.1021/bm060849t
[[VPGVG][VPGHG]4]8	64.2	207	25	35920	10.1021/nl300630c
[[VPGVG][VPGHG]4]16	25.7	407	25	44718	10.1021/nl300630c
[[VPGVG][VPGHG]4]20	26.2	507	25	8941	10.1021/nl300630c
[[VPGVG][VPGHG]2[VPGGG][VPGAG]]4	50.9	107	100	17155	10.1021/bm100571j
[[VPGVG][VPGHG]2[VPGGG][VPGAG]]8	35.5	207	100	25368	10.1021/bm100571j
[[VPGVG][VPGHG]2[VPGGG][VPGAG]]12	27.9	307	100	50008	10.1021/bm100571j
[[VPGVG][VPGHG]2[VPGGG][VPGAG]]24	23.6	607	100	25589	10.1021/bm100571j
[[VPGVG][VPGGG]]32	71.17	327	25	34743	10.1021/nl300630c
[[VPGVG][VPGIG]3[VPGEG]]16	57.68405	408	100	68655	10.1021/bm100571j
[[VPGVG][VPGIG]3[VPGEG]]32	50.13177	808	100	50540	10.1021/bm100571j
[[VPGVG]6[VPGLG]4]12	16.3	607	50	49676	10.1038/s41551-017-0078
[VPGG]160	51.7	640	800	50408	10.1002/bip.360310502
[VPGVGVPGG]70	44.9	630	600	51919	10.1002/bip.360310502/
[IPGG]160	29.3	640	400	50840	10.1002/bip.360310502/
[LPGVG]120	16.2	600	200	50840	10.1002/bip.360310502/
[VPAVG]120	34.2	600	200	103205	10.1002/bip.360310502/
[[VPGVG]0.77[VPGIG]0.23]250	23.8	1250	400	103654	10.1002/bip.360320913
[[VPGVG]0.64[VPGIG]0.36]250	22.3	1250	400	102953	10.1002/bip.360320913

[[VPGVG]0.84[VPGLG]0.16]250	22.8	1250	400	103485	10.1002/bip.360320913
[[VPGVG]0.69[VPGLG]0.31]250	20.4	1250	400	116485	10.1002/bip.360320913
[VPGIG]250	14.1	1375	400	105897	10.1002/bip.360320914
[IPGVG]250	11.2	1250	400	105897	10.1002/bip.360320914
[LPGVG]250	15.4	1250	400	110134	10.1002/bip.360320914
[[VPGVG]0.45[IPGVG]0.55]250	18.9	1250	400	109012	10.1002/bip.360320914
[[VPGVG]0.53[LPGVG]0.47]250	21	1250	400	11923	10.1002/bip.360320914
[VTPAVG]20	LCST too high	132	50	21214	10.1038/nmat4418
[LGVPAG]40	31.9	252	50	27970	US20120121709A1
[TVPGAG]55	LCST too high	342	50	18595	10.1038/nmat4418
[APGVG]45	LCST too high	237	50	14146	10.1038/nmat4418
[VRPVG]25	LCST too high	137	50	11042	10.1038/nmat4418
[VAPGVG]20	LCST too high	132	50	14614	10.1038/nmat4418
[VAPDG]30	LCST too high	162	50	27238	10.1038/nmat4418
[VPSDDYGVG]29	LCST too high	273	50	19154	10.1038/nmat4418
[VPSALYGVG]21	LCST too low	201	50	12540	10.1038/nmat4418
[GAPFGFAIPMGAGFPTGGLAPFGMGLPAGM]4	LCST too low	132	50	16421	US20120121709A1
[GVLPGVG]28	33.3	199	25	30883	10.3109/1061186X.2015.1077847
[GAGVPG]70	LCST too high	423	25	15179	10.3109/1061186X.2015.1077847
[VPGFGAGAG]21	35.7	192	25	65448	10.3109/1061186X.2015.1077847
[VPGLGAGAG]96	38.6	867	25	25538	10.3109/1061186X.2015.1077847

References

1. V. Aseyev, H. Tenhu, F. M. Winnik, (Springer, Berlin, Heidelberg, 2010; http://link.springer.com/10.1007/12_2010_57), pp. 29–89.

2. D. Chandler, Interfaces and the driving force of hydrophobic assembly. *Nature*. **437**, 640–647 (2005).

3. I. Saha, U. Maulik, S. Bandyopadhyay, D. Plewczynski, Fuzzy clustering of physicochemical and biochemical properties of amino acids. *Amino Acids*. **43**, 583–94 (2012).

4. S. Simm, J. Einloft, O. Mirus, E. Schleiff, 50 years of amino acid hydrophobicity scales: revisiting the capacity for peptide classification. *Biol. Res.* **49**, 31 (2016).

5. J. C. Salgado, I. Rapaport, J. A. Asenjo, Is it possible to predict the average surface hydrophobicity of a protein using only its amino acid composition? *J. Chromatogr.* **1075**, 133–143 (2005).

6. F. G. Quiroz, A. Chilkoti, Sequence heuristics to encode phase behaviour in intrinsically disordered protein polymers. *Nat. Mater.* **advance on** (2015), doi:10.1038/nmat4418.

7. S. Rauscher, S. Baud, M. Miao, F. W. Keeley, R. Pomès, Proline and Glycine Control Protein Self-Organization into Elastomeric or Amyloid Fibrils. *Structure*. **14**, 1667–1676 (2006).

8. S. Rauscher, R. Pomès, The liquid structure of elastin. *Elife*. **6** (2017), doi:10.7554/eLife.26526.

9. N. K. Li, F. G. Quiroz, C. K. Hall, A. Chilkoti, Y. G. Yingling, Molecular Description of the LCST Behavior of an Elastin-Like Polypeptide. *Biomacromolecules*. **15**, 3522–3530 (2014).

10. A. M. Tamburro, B. Bochicchio, A. Pepe, Dissection of Human Tropoelastin: Exon-By-Exon Chemical Synthesis and Related Conformational Studies. *Biochemistry*. **42**, 13347–13362 (2003).

11. L. D. Muiznieks, A. S. Weiss, F. W. Keeley, Structural disorder and dynamics of elastin. *Biochem. Cell Biol.* **88**, 239–50 (2010).

12. H. G. Schild, Poly(N-isopropylacrylamide): experiment, theory and application.

Prog. Polym. Sci. **17**, 163–249 (1992).

13. Dan W. Urry, *What Sustains Life?: Consilient Mechanisms for Protein-Based Machines and Materials* (https://books.google.com/books?id=KNEmcjxANegC&pg=PA576&lpg=PA576&dq=urry+phase+diagram&source=bl&ots=ibun141iLp&sig=znQ1R3MsTwPkN5ZtgMDirdnSHps&hl=en&sa=X&ved=2ahUKEwiL-LD75bLfAhWQct8KHbBBDPQQ6AEwBnoECAgQAQ#v=onepage&q&f=false).

14. D. W. Urry, The change in Gibbs free energy for hydrophobic association. *Chem. Phys. Lett.* **399**, 177–183 (2004).

15. D. E. Meyer, A. Chilkoti, Quantification of the Effects of Chain Length and Concentration on the Thermal Behavior of Elastin-like Polypeptides. *Biomacromolecules*. **5**, 846–851 (2004).

16. D. E. Meyer, a Chilkoti, Purification of recombinant proteins by fusion with thermally-responsive polypeptides. *Nat. Biotechnol.* **17**, 1112–5 (1999).

17. J. A. Mackay, D. J. Callahan, K. N. Fitzgerald, A. Chilkoti, Quantitative Model of the Phase Behavior of Recombinant pH-Responsive Elastin-Like Polypeptides, 2873–2879 (2010).

18. M. Rubinstein, R. Colby, (2003), vol. 23.

19. S. Fluegel, K. Fischer, J. R. McDaniel, A. Chilkoti, M. Schmidt, Chain stiffness of elastin-like polypeptides. *Biomacromolecules*. **11**, 3216–8 (2010).

20. J. R. McDaniel, D. C. Radford, A. Chilkoti, A unified model for de novo design of elastin-like polypeptides with tunable inverse transition temperatures. *Biomacromolecules*. **14**, 2866–72 (2013).

21. N. K. Adam, The Physics and Chemistry of Surfaces.

22. A. Ferguson, Relations between thermo-physical properties. *Proc. Phys. Soc.* **52**, 759–763 (1940).

23. B. Widom, Equation of State in the Neighborhood of the Critical Point. *J. Chem. Phys.* **43**, 3898–3905 (1965).

24. F. J. Blas *et al.*, Universal scaling behaviour of surface tension of molecular chains. *Artic. J. Chem. Phys.* **137**, 24702 (2012).

25. G. Galliero, Surface tension of short flexible Lennard-Jones chains: Corresponding states behavior. *J. Chem. Phys.* **133**, 074705 (2010).

26. W. Hassouneh, E. B. Zhulina, A. Chilkoti, M. Rubinstein, Elastin-like Polypeptide Diblock Copolymers Self-Assemble into Weak Micelles. *Macromolecules*. **48**, 4183–4195 (2015).

27. P. Shi *et al.*, Elastin-based protein polymer nanoparticles carrying drug at both corona and core suppress tumor growth in vivo. *J. Control. Release*. **171**, 330–338 (2013).

28. M. Shah *et al.*, Biodegradation of elastin-like polypeptide nanoparticles. *Protein Sci.* **21**, 743–750 (2012).

29. K. Trabbic-Carlson, L. A. Setton, A. Chilkoti, Swelling and Mechanical Behaviors of Chemically Cross-Linked Hydrogels of Elastin-like Polypeptides. *Biomacromolecules*. **4**, 572–580 (2003).

30. M. K. McHale, L. A. Setton, A. Chilkoti, Synthesis and in Vitro Evaluation of Enzymatically Cross-Linked Elastin-Like Polypeptide Gels for Cartilaginous Tissue Repair. *Tissue Eng.* **11**, 1768–1779 (2005).

31. D. W. Lim, K. Trabbic-Carlson, J. A. MacKay, A. Chilkoti, Improved Non-chromatographic Purification of a Recombinant Protein by Cationic Elastin-like Polypeptides. *Biomacromolecules*. **8**, 1417–1424 (2007).

32. D. J. Callahan *et al.*, Triple Stimulus-Responsive Polypeptide Nanoparticles That Enhance Intratumoral Spatial Distribution. *Nano Lett.* **12**, 2165–2170 (2012).

33. T. Yamaoka *et al.*, Mechanism for the Phase Transition of a Genetically Engineered Elastin Model Peptide (VPGIG) 40 in Aqueous Solution. *Biomacromolecules*. **4**, 1680–1685 (2003).

34. C.-H. Luan, T. M. Parker, K. U. Prasad, D. W. Urry, Differential scanning calorimetry studies of NaCl effect on the inverse temperature transition of some elastin-based polytetra-, polypenta-, and polynonapeptides. *Biopolymers*. **31**, 465–475 (1991).

35. C.-H. Luan, T. M. Parker, D. C. Gowda, D. W. Urry, Hydrophobicity of amino acid residues: Differential scanning calorimetry and synthesis of the aromatic analogues of the polypentapeptide of elastin. *Biopolymers*. **32**, 1251–1261 (1992).

36. S. Cho, S. Dong, K. N. Parent, M. Chen, Immune-tolerant elastin-like polypeptides (iTEPs) and their application as CTL vaccine carriers. *J. Drug Target.* **24**, 328–339 (2016).

37. D. W. Urry *et al.*, Hydrophobicity scale for proteins based on inverse temperature transitions. *Biopolymers*. **32**, 1243–1250 (1992).

38. D. E. Meyer, A. Chilkoti, Genetically Encoded Synthesis of Protein-Based Polymers with Precisely Specified Molecular Weight and Sequence by Recursive Directional Ligation: Examples from the Elastin-like Polypeptide System. *Biomacromolecules*. **3**, 357–367 (2002).

39. J. A. MacKay, D. J. Callahan, K. N. FitzGerald, A. Chilkoti, Quantitative Model of the Phase Behavior of Recombinant pH-Responsive Elastin-Like Polypeptides. *Biomacromolecules*. **11**, 2873–2879 (2010).

40. N. C. Tang, A. Chilkoti, Combinatorial codon scrambling enables scalable gene synthesis and amplification of repetitive proteins. *Nat. Mater.* **15**, 419–424 (2016).

41. R. Carlson, Time for New DNA Synthesis and Sequencing Cost Curves (2014), (available at http://www.synthesis.cc/2014/02/time-for-new-cost-curves-2014.html).

42. S. Mueller, J. R. Coleman, E. Wimmer, Putting synthesis into biology: a viral view of genetic engineering through de novo gene and genome synthesis. *Chem. Biol.* **16**, 337–47 (2009).

43. S. Kosuri *et al.*, Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat. Biotechnol.* **28**, 1295–9 (2010).

44. M. Goldberg, BioFab: Applying Moore's Law to DNA Synthesis. *Ind. Biotechnol.* **9**, 10–12 (2013).

45. S. Kosuri, G. M. Church, Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods*. **11**, 499–507 (2014).

46. K. Jayaraman, J. Shah, J. Fyles, PCR mediated gene synthesis. *Nucleic Acids Res.* **17**, 4403 (1989).

47. H. O. Smith, C. A. Hutchison, C. Pfannkoch, J. C. Venter, Generating a synthetic genome by whole genome assembly: φX174 bacteriophage from synthetic oligonucleotides. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 15440 (2003).

48. M. Gendreau, J.-Y. Potvin, *Handbook of Metaheuristics* (Springer, ed. 2, 2010; http://www.springer.com/us/book/9781441916631), vol. 146.

49. P. A. Pevzner, H. Tang, M. S. Waterman, An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 9748–53 (2001).

50. W. R. Whitaker, H. Lee, A. P. Arkin, J. E. Dueber, Avoidance of truncated proteins from unintended ribosome binding sites within heterologous protein coding sequences. *ACS Synth. Biol.* **4**, 249–57 (2015).

51. G. Laporte, The traveling salesman problem: An overview of exact and approximate algorithms. *Eur. J. Oper. Res.* **59**, 231–247 (1992).

52. R. Watrigant, M. Bougeret, R. Giroudeau, The k-Sparsest Subgraph Problem (2012).

53. C. Miller, A. Tucker, R. Zemlin, Integer programming formulation of traveling salesman problems. *J. ACM*, 326–329 (1960).

54. M. Desrochers, G. Laporte, Improvements and extensions to the Miller-Tucker-Zemlin subtour elimination constraints. *Oper. Res. Lett.* **10**, 27–36 (1991).

55. D. Gusfield, Algorithms on Strings, Trees, and Sequences. *ACM SIGACT News*. **28** (1997), pp. 41–60.

56. J. R. McDaniel, J. A. Mackay, F. G. Quiroz, A. Chilkoti, Recursive directional ligation by plasmid reconstruction allows rapid and seamless cloning of oligomeric genes. *Biomacromolecules*. **11**, 944–52 (2010).

57. D. Anderson, K. Maugh, Escherichia coli expression vector encoding bioadhesive precursor protein analogs comprising three to twenty repeats of the decapeptide (Ala-Lys-Pro-Ser-Tyr-Pro-. *US Pat. 5,149,657* (1992) (available at http://www.google.com/patents/US5149657).

58. R. E. Lyons *et al.*, Design and facile production of recombinant resilin-like polypeptides: gene construction and a rapid protein purification method. *Protein Eng. Des. Sel.* **20**, 25–32 (2007).

59. R. S.-C. Su, J. N. Renner, J. C. Liu, Synthesis and characterization of recombinant abductin-based proteins. *Biomacromolecules*. **14**, 4301–8 (2013).

60. J. Cappello, F. Ferrari, C. Richardson, Methods for preparing synthetic repetitive

DNA. *US Pat. 5,641,648* (1997) (available at http://www.google.com/patents/US5641648).

61. J. Cappello, S. Causey, Peptides comprising repetitive units of amino acids and DNA sequences encoding the same. *US Pat. 6,018,030*. **262** (2000) (available at http://www.google.com/patents/US6018030).

62. M. Amiram, F. Quiroz, D. Callahan, A. Chilkoti, A highly parallel method for synthesizing DNA repeats enables the discovery of "smart" protein polymers. *Nat. Mater.* **10**, 141–8 (2011).

63. R. S. Farmer, A. Top, L. M. Argust, S. Liu, K. L. Kiick, Evaluation of conformation and association behavior of multivalent alanine-rich polypeptides. *Pharm. Res.* **25**, 700–8 (2008).

64. O. Shur, S. Banta, Rearranging and concatenating a native RTX domain to understand sequence modularity. *Protein Eng. Des. Sel.* **26**, 171–80 (2013).

65. D. Steiner, P. Forrer, A. Plückthun, Efficient selection of DARPins with sub-nanomolar affinities using SRP phage display. *J. Mol. Biol.* **382**, 1211–27 (2008).

66. B. W. Lee *et al.*, Strongly binding cell-adhesive polypeptides of programmable valencies. *Angew. Chemie - Int. Ed.* **49**, 1971–1975 (2010).

67. B. Polypeptides *et al.*, Design and Preparation of -Sheet Forming Repetitive and. *Biomacromolecules*, 1487–1497 (2007).

68. W. Petka, J. Harden, K. McGrath, D. Wirtz, D. Tirrell, Reversible hydrogels from self-assembling artificial proteins. *Science (80-.).* **281**, 389–392 (1998).

69. N. E. Davis, S. Ding, R. E. Forster, D. M. Pinkas, A. E. Barron, Modular enzymatically crosslinked protein polymer hydrogels for in situ gelation. *Biomaterials*. **31**, 7288–97 (2010).

70. K. Elmorjani *et al.*, Synthetic genes specifying periodic polymers modelled on the repetitive domain of wheat gliadins: conception and expression. *Biochem. Biophys. Res. Commun.* **239**, 240–246 (1997).

71. D. E. Meyer, K. Trabbic-Carlson, A. Chilkoti, Protein purification by fusion with an environmentally responsive elastin-like polypeptide: Effect of polypeptide length on the purification of thioredoxin. *Biotechnol. Prog.* **17**, 720–728 (2001).

72. J. P. O'Brien *et al.*, in *Silk Polymers: Materials Science and Biotechnology* (1994; http://www.beta-sheet.org/resources/77-Adeno-Fiber-ACSSymp1993.pdf), vol. 544, pp. 104–117.

73. T. Cermak *et al.*, Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res.* **39**, e82 (2011).

74. D. M. Widmaier *et al.*, Engineering the Salmonella type III secretion system to export spider silk monomers. *Mol. Syst. Biol.* **5**, 309 (2009).

75. D. G. Ousterout *et al.*, Reading frame correction by targeted genome editing restores dystrophin expression in cells from Duchenne muscular dystrophy patients. *Mol. Ther.* **21**, 1718–26 (2013).

76. C. Engler, R. Gruetzner, R. Kandzia, S. Marillonnet, Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. *PLoS One*. **4**, e5553 (2009).

77. D. E. Meyer, A. Chilkoti, Genetically encoded synthesis of protein-based polymers with precisely specified molecular weight and sequence by recursive directional ligation: examples from the elastin-like polypeptide system. *Biomacromolecules*. **3**, 357–67 (2002).

78. D. G. Gibson *et al.*, Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*. **6**, 343–345 (2009).

79. T. Tuller, Y. Y. Waldman, M. Kupiec, E. Ruppin, Translation efficiency is determined by both codon bias and folding energy. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 3645–50 (2010).

80. M. Welch, A. Villalobos, C. Gustafsson, J. Minshull, You're one in a googol: optimizing genes for protein expression. *J. R. Soc. Interface*. **6 Suppl 4**, S467-76 (2009).

81. G. Kudla, A. W. Murray, D. Tollervey, J. B. Plotkin, Coding-sequence determinants of gene expression in Escherichia coli. *Science*. **324**, 255–8 (2009).

82. D. B. Goodman, G. M. Church, S. Kosuri, Causes and effects of N-terminal codon bias in bacterial genes. *Science*. **342**, 475–9 (2013).

83. S. M. Richardson, S. J. Wheelan, R. M. Yarrington, J. D. Boeke, GeneDesign: rapid, automated design of multikilobase synthetic genes. *Genome Res.* **16**, 550–6 (2006).

84. A. Blanchard, R. Kaiser, L. Hood, High-density oligonucleotide arrays. *Biosens. Bioelectron.* **11**, 687–690 (1996).

85. D. A. Lashkari, S. P. Hunicke-Smith, R. M. Norgren, R. W. Davis, T. Brennan, An automated multiplex oligonucleotide synthesizer: development of high-throughput, low-cost DNA synthesis. *Proc. Natl. Acad. Sci.* **92**, 7912 (1995).

86. a Ellington, J. D. Pollard, *Curr. Protoc. Nucleic Acid Chem.*, in press, doi:10.1002/0471142700.nca03cs00.

87. X. Zhou *et al.*, Microfluidic PicoArray synthesis of oligodeoxynucleotides and simultaneous assembling of multiple DNA sequences. *Nucleic Acids Res.* **32**, 5409–17 (2004).

88. F. B. Dean, J. R. Nelson, T. L. Giesler, R. S. Lasken, Rapid Amplification of Plasmid and Phage DNA Using Phi29 DNA Polymerase and Multiply-Primed Rolling Circle Amplification. *Genome Res.* **11**, 1095 (2001).

89. C. M. Hommelsheim, L. Frantzeskakis, M. Huang, B. Ülker, PCR amplification of repetitive DNA: a limitation to genome editing technologies and many other applications. *Sci. Rep.* **4**, 5052 (2014).

90. K.-K. Wong, L. M. Markillie, J. D. Saffer, A novel method for producing partial restriction digestion of DNA fragments by PCR with 5-methyl-CTP. *Nucleic Acids Res.* **25**, 4169–4171 (1997).

91. J. J. Schwartz, C. Lee, J. Shendure, Accurate gene synthesis with tag-directed retrieval of sequence-verified DNA molecules. *Nat. Methods*. **9**, 913–915 (2012).

92. D. Johnston *et al.*, Illumina MiSeq 16S amplicon sequence analysis of bovine respiratory disease associated bacteria in lung and mediastinal lymph node tissue. *BMC Vet. Res.* **13**, 118 (2017).

93. B. Kerr, M. a Riley, M. W. Feldman, B. J. M. Bohannan, Local dispersal promotes biodiversity in a real-life game of rock-paper-scissors. *Nature*. **418**, 171–4 (2002).

94. S.-L. Dahlroth, P. Nordlund, T. Cornvik, Colony filtration blotting for screening soluble expression in Escherichia coli. *Nat. Protoc.* **1**, 253–258 (2006).

95. C.-T. Lin *et al.*, Automated purification of recombinant proteins: combining high-throughput with high yield. *Protein Expr. Purif.* **47**, 16–24 (2006).

96. W. A. Duetz, Microtiter plates as mini-bioreactors: miniaturization of fermentation methods. *Trends Microbiol.* **15**, 469–475 (2007).

97. F. Schäfer *et al.*, Automated High-Throughput Purification of 6xHis-Tagged Proteins. *J. Biomol. Tech.* **13**, 131–142.

98. J. Koehn, I. Hunt, High Throughput Protein Expression and Purification. **498** (2009), doi:10.1007/978-1-59745-196-3.

99. T. Christensen, K. Trabbic-Carlson, W. Liu, A. Chilkoti, Purification of recombinant proteins from Escherichia coli at low expression levels by inverse transition cycling. *Anal. Biochem.* **360**, 166–8 (2007).

100. A. Kaul, in *Aqueous Two-Phase Systems* (Humana Press, New Jersey, 2000; http://link.springer.com/10.1385/1-59259-028-4:11), pp. 11–21.

101. J. H. WAN, Investigation into the phase separation behavior of concentrated elastin-like polypeptide solutions (2016), (available at http://rave.ohiolink.edu/etdc/view?acc_num=csu1471923273).

102. Y. Cho *et al.*, Effects of Hofmeister anions on the phase transition temperature of elastin-like polypeptides. *J. Phys. Chem. B*. **112**, 13765–71 (2008).

103. S. R. MacEwan, W. Hassouneh, A. Chilkoti, Non-chromatographic purification of recombinant elastin-like polypeptides and their fusions with peptides and proteins from Escherichia coli. *J. Vis. Exp.*, e51583 (2014).

104. N. J. Greenfield, Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions. *Nat. Protoc.* **1**, 2527–35 (2006).

105. Y.-B. Yan, J. Zhang, H.-W. He, H.-M. Zhou, Oligomerization and aggregation of bovine pancreatic ribonuclease A: characteristic events observed by FTIR spectroscopy. *Biophys. J.* **90**, 2525–33 (2006).

106. L. D. Muiznieks, F. W. Keeley, Proline periodicity modulates the self-assembly properties of elastin-like polypeptides. *J. Biol. Chem.* **285**, 39779–89 (2010).

107. U. Nobbmann *et al.*, Dynamic light scattering as a relative tool for assessing the molecular integrity and stability of monoclonal antibodies. *Biotechnol. Genet. Eng. Rev.* **24**, 117–128 (2007).

108. K. Phillips, A. H. de la Peña, The combined use of the thermofluor assay and thermoQ analytical software for the determination of protein stability and buffer optimization as an aid in protein crystallization. *Curr. Protoc. Mol. Biol.* **Chapter 10**, 1–15 (2011).

109. K. Shiba, Natural and artificial peptide motifs: their origins and the application of motif-programming. *Chem. Soc. Rev.* **39**, 117–26 (2010).

110. H. Saito, T. Minamisawa, K. Shiba, Motif programming: a microgene-based method for creating synthetic proteins containing multiple functional motifs. *Nucleic Acids Res.* **35**, e38–e38 (2007).

111. S. Roberts *et al.*, Injectable tissue integrating networks from recombinant polypeptides with tunable order. *Nat. Mater.* **17**, 1154–1163 (2018).

112. W. Hassouneh, S. R. MacEwan, A. Chilkoti, in *Methods in enzymology* (2012; http://www.ncbi.nlm.nih.gov/pubmed/22208987), vol. 502, pp. 215–237.

113. O. G. Troyanskaya, O. Arbell, Y. Koren, G. M. Landau, A. Bolshoy, Sequence complexity profiles of prokaryotic genomic sequences: A fast algorithm for calculating linguistic complexity. *Bioinformatics*. **18**, 679–688 (2002).

114. A. S. Holehouse, R. K. Das, J. N. Ahad, M. O. G. Richardson, R. V. Pappu, CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophys. J.* **112**, 16–21 (2017).

115. R. Wuttke *et al.*, Temperature-dependent solvation modulates the dimensions of disordered proteins. *Proc. Natl. Acad. Sci.* **111**, 5213–5218 (2014).

116. A. G. Kikhney, D. I. Svergun, A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett.* **589**, 2570–2577 (2015).

117. J. M. R. Baker *et al.*, CFTR regulatory region interacts with NBD1 predominantly via multiple transient helices. *Nat. Struct. Mol. Biol.* **14**, 738–745 (2007).

118. D. I. Pretto *et al.*, Structural Dynamics and Single-Stranded DNA Binding Activity of the Three N-Terminal Domains of the Large Subunit of Replication Protein A from Small Angle X-ray Scattering. *Biochemistry*. **49**, 2880–2889 (2010).

119. J. Heyda, J. Dzubiella, Thermodynamic Description of Hofmeister Effects on the LCST of Thermosensitive Polymers. *J. Phys. Chem. B*. **118**, 10979–10988 (2014).

120. J. Heyda *et al.*, Guanidinium can both Cause and Prevent the Hydrophobic Collapse of Biomacromolecules. *J. Am. Chem. Soc.* **139**, 863–870 (2017).

121. Y. Zhang *et al.*, Effects of Hofmeister Anions on the LCST of PNIPAM as a Function of Molecular Weight. *J. Phys. Chem. C*. **111**, 8916–8924 (2007).

122. G. Longo, M. Montévil, From physics to biology by extending criticality and symmetry breakings. *Prog. Biophys. Mol. Biol.* **106**, 340–347 (2011).

123. T. Mora, W. Bialek, Are Biological Systems Poised at Criticality? *J. Stat. Phys.* **144**, 268–302 (2011).

124. T. Christensen, W. Hassouneh, K. Trabbic-Carlson, A. Chilkoti, Predicting transition temperatures of elastin-like polypeptide fusion proteins. *Biomacromolecules*. **14**, 1514–9 (2013).

125. D. Chandler, Interfaces and the driving force of hydrophobic assembly. *Nature*. **437**, 640–647 (2005).

126. J. L. Ross, The Dark Matter of Biology. *Biophys. J.* **111**, 909–916 (2016).

127. M. Rubinstein, Polymer physics-The ugly duckling story: Will polymer physics ever become a part of "proper" physics? *J. Polym. Sci. Part B Polym. Phys.* **48**, 2548–2551 (2010).

Biography

Nicholas Tang graduated from Duke University in May of 2011 with a B.S.E. in Biomedical Engineering and a B.S.E. in Electrical and Computer Engineering. He pursued a Ph.D. in the Department of Biomedical Engineering at Duke University, where he was presented with an NIH T32 fellowship for predoctoral training in the Duke Center for Biomolecular and Tissue Engineering (CBTE) in 2012 and the Kewaunee Senior Graduate Poster Session Award in 2017. His work has helped produce the following publications:

1. **N. C. Tang** & A. Chilkoti (2016). Combinatorial codon scrambling enables scalable gene synthesis and amplification of repetitive proteins. ***Nature Materials***, *15*(4), 419–24.

2. J. Quan, I. Saaem, **N. Tang**, S. Ma, N. Negre, H. Gong, K. White, J. Tian (2011). Parallel on-chip gene synthesis and application to optimization of protein expression. ***Nature Biotechnology***, *29*(5), 449–52.

3. **N. Tang**, S. Ma, & J. Tian (2013). New tools for cost-effective DNA synthesis. **In *Synthetic Biology*** (Vol. 1, pp. 3–21). Elsevier Inc.

4. S. Ma, **N. Tang**, & J. Tian (2012). DNA synthesis, assembly and applications in synthetic biology. ***Current Opinion in Chemical Biology***, *16*(3-4), 260–7.

5. F. Reza, **N. Tang**, … J. Tian (2007). Engineering novel synthetic biological systems. ***Synthetic Biology, IET***, *1*(1.2), 48–52.