Protein Engineering for Biomedicine and Beyond

Jennifer Phipps McCord

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Chemistry

Tijana Z. Grove, Chair
Felicia A. Etzkorn
Alan R. Esker
Mark E. Van Dyke

May 8th, 2019
Blacksburg, VA

Keywords: protein engineering, consensus design, repeat proteins, cellulose binding module, keratin

Protein Engineering for Biomedicine and Beyond

Jennifer Phipps McCord

ABSTRACT

Many applications in biomedicine, research, and industry require recognition agents with specificity and selectivity for their target. Protein engineering enables the design of scaffolds that can bind targets of interest while increasing their stability, and expanding the scope of applications in which these scaffolds will be useful.

Repeat proteins are instrumental in a wide variety of biological processes, including the recognition of pathogen-associated molecular patterns by the immune system. A number of successes using alternative immune system repeat protein scaffolds have expanded the scope of recognition agents available for targeting glycans and glycoproteins in particular. We have analyzed the innate immune genes of a freshwater polyp and found that they contained particularly long contiguous domains with high sequence similarity between repeats in these domains. We undertook statistical design to create a binding protein based on the *H. magnipapillata* innate immune TPR proteins.

My second research project focused on creating a protein to bind cellulose, as it is the most abundant and inexpensive source of biomass and therefore is widely considered a possible source for liquid fuel. However, processing costs have kept lignocellulosic fuels from competing commercially with starch-based biofuels. In recent years a strategy to protect processing enzymes with synergistic proteins emerged to reduce the amount of enzyme necessary for lignocellulosic biofuel production. Simultaneously, protein engineering approaches have been developed to optimize proteins for function and

stability enabling the use of proteins under non-native conditions and the unique conditions required for any necessary application. We designed a consensus protein based on the carbohydrate-binding protein domain CBM1 that will bind to cellulosic materials. The resulting designed protein is a stable monomeric protein that binds to both microcrystalline cellulose and amorphous regenerated cellulose thin films. By studying small changes to the binding site, we can better understand how these proteins bind to different cellulose-based materials in nature and how to apply their use to industrial applications such as enhancing the saccharification of lignocellulosic feedstock for biofuel production.

Biomaterials made from natural human hair keratin have mechanical and biochemical properties that make them ideal scaffolds for tissue engineering and wound healing. However, the extraction process leads to protein degradation and brings with it byproducts from hair, which can cause unfavorable immune responses. Recombinant keratin biomaterials are free from these disadvantages, while heterologous expression of these proteins allows us to manipulate the primary sequence. We endeavored to add an RGD sequence to facilitate cell adhesion to the recombinant keratin proteins, to demonstrate an example of useful sequence modification.

Protein Engineering for Biomedicine and Beyond

Jennifer Phipps McCord

GENERAL AUDIENCE ABSTRACT

Many applications in medicine and research require molecular sensors that bind their target tightly and selectively, even in complex mixtures. Mammalian antibodies are the best-studied examples of these sensors, but problems with the stability, expense, and selectivity of these antibodies have led to the development of alternatives.

In the search for better sensors, repeat proteins have emerged as one promising class, as repeat proteins are relatively simple to design while being able to bind specifically and selectively to their targets. However, a drawback of commonly used designed repeat proteins is that their targets are typically restricted to proteins, while many targets of biomedical interest are sugars, such as those that are responsible for blood types. Repeat proteins from the immune system, on the other hand, bind targets of many different types. We looked at the unusual immune system of a freshwater polyp as inspiration to design a new repeat protein to recognize nonprotein targets.

My second research project focused on binding cellulose, as it is the most abundant and inexpensive source of biological matter and therefore is widely considered a possible source for liquid fuel. However, processing costs have kept cellulose-based fuels from competing commercially with biofuel made from corn and other starchy plants. One strategy to lower costs relies on using helper proteins to reduce the amount of enzyme needed to break down the cellulose, as enzymes are the most expensive part of processing. We designed such a protein for this function to be more stable than natural

proteins currently used. The resulting designed protein binds to multiple cellulose structures. Designing a protein from scratch also allows us to study small changes to the binding site, allowing us to better understand how these proteins bind to different cellulose-based materials in nature and how to apply their use to industrial applications.

Biomaterials made from natural human hair keratin have mechanical and biochemical properties that make them ideal for tissue engineering and wound healing applications. However, the process by which these proteins are extracted from hair leads to some protein degradation and brings with it byproducts from hair, which can cause unfavorable immune responses. Making these proteins synthetically allows us to have pure starting material, and lets us add new features to the proteins, which translates into materials better tailored for their applications. We discuss here one example, in which we added a cell-binding motif to a keratin protein sequence.

Acknowledgments

I am incredibly grateful to my advisor, Dr. Tijana Grove, for her mentorship throughout my time at Virginia Tech. Her expertise, constructive feedback, willingness to entertain strange ideas and innumerable drafts have facilitated my growth as a scientist, as a teacher, and as a writer. Being in her group has undoubtedly helped me become a better mentor myself, and I am forever grateful for her guidance.

I would like to thank Dr. Etzkorn for her generosity in volunteering her group meetings to help me practice public speaking (and Leanne and Paul for listening), for sharing her chemicals, for her willingness to share her experiences with the Graduate Women in Chemistry meetings, and all the other help she has given me over my graduate school years.

Dr. Esker and Dr. Van Dyke have been incredibly supportive in our collaborations, and I am grateful for their encouragement and assistance with experiments and data. I would also like to thank their students Jianzhao Liu and Alexis Trent for their help and willingness to teach me new skills.

My colleagues in the department have been excellent mentors and friends throughout my time in graduate school. I'd like to thank Christina Kim for her friendship, willingness to listen to my presentations before they make sense, for being a great sounding board for my research, and most especially for agreeing to swap unfavorable lab chores. I'm grateful to Kristina Roth and Rachael Parker for their mentorship in helping my work get off the ground and improve, for their eternal willingness to discuss the day's Highlands seminar over chips and salsa, to Grey Fritz for being a fantastic undergraduate in the lab, and to the entire Grove lab for helping with ideas, polishing

research, and celebrations. Thanks to Ashley Gates for her help proofreading emails, willingness to commiserate when the world gets us down, being an excellent office buddy and good friend. I'd also like to thank Leanne Aakjar, Paul Acoria, and Justin Grams for the frequent laughter in the office and for only occasionally getting me caught in paper plane crossfire. Thanks to Katie (and Zach) Heifferon for friendship, exercise, game nights, encouragement to leave the house, and allowing Pip to terrorize your cats with attempts to play.

Thanks to the Blacksburg Master Chorale for giving the other half of my brain an outlet on Tuesday nights. I am thankful to everyone in the group for all their support, laughter, and willingness to help in any way that they can. I'd especially like to thank Elizabeth Cox, Deb Call, and the rest of the soprano section for their friendship.

Most importantly, I'd like to thank my family for their support. Mac and Judy McCord's generosity, excellent food, company, and pet therapy have been instrumental in my ability to stay sane in graduate school. Thanks to my parents for their support, willingness to read papers and give helpful advice. I'd also like to thank Betsy McCord for her inclination to edit dissertation chapters while sailing back from the Bahamas. Thanks to Drew, an occasional supplier of tasty food but a consistent supplier of sanity and reminder that work isn't everything.

I will be forever grateful to everyone who has supported me along this journey. Thank you.

Attributions

Dr. Tijana Z. Grove is a professor in the Department of Chemistry. She is the author's research advisor and mentor.

Dr. Alan R. Esker is a professor in the Department of Chemistry and a member of the author's PhD committee. He was a collaborator on Chapter 5.

V. Grey Fritz is an undergraduate student in Dr. Grove's research group and was mentored by the author. She was a contributor on Chapters 4 and 5.

Jianzhao Liu is a PhD graduate student in Dr. Esker's research group and was a contributor on Chapter 5.

Dr. Mark R. Van Dyke is a professor in the Department of Biomedical Engineering and Mechanics and a member of the author's PhD committee. He was a collaborator on Chapter 6.

Table of Contents

Chapter 1. Engineered Proteins

1.1 Dissertation overview

This dissertation describes protein engineering approaches to create new molecular recognition agents for biomedicine and industrial applications. An underlying theme in our work is the expansion of the scope of molecular recognition agents for glycan and glycoprotein ligands, as current methods to do so are not specific or selective for their target.

1.2 Engineering Non-Immunoglobulin Immune System Proteins

Chapter 3 discusses recent efforts in the field of protein engineering to take advantage of non-immunoglobulin proteins of the immune system. A drawback of many well-characterized non-immunoglobulin binding proteins is that they are limited to protein and peptide targets. Pathogen- and damage-associated molecular patterns indicative of danger to the organism run the gamut from nucleic acids to glycoproteins to small molecules, and the innate immune system recognizes this chemical diversity. Designed alternative immune proteins have been able to take advantage of this natural binding diversity to recognize glycan and glycoprotein ligands in addition to peptides and proteins. This chapter presents an overview of these scaffolds, along with successes in noncognate ligand recognition.

1.3 Repeat Domains of the *Hydra Magnipapillata* Innate Immune System

Recent phylogenetic analyses of the protein domains associated with the innate immune system found that in early diverging organisms, the Nucleotide Binding (NB-ARC) domain associated with innate immune signaling was paired with unusual recognition domains. *Hydra magnipapillata* had a particularly large repertoire of unusual recognition domains, and we

therefore undertook the challenge of using these as a basis for the design of a new recognition scaffold. Chapter 4 details our efforts to make a sensor based on these recognition domains to bind physiologically relevant targets, particularly glycoproteins.

1.4 Consensus Design of a Family 1 Cellulose Binding Module

Polysaccharide targets are also of interest to industrial applications. Lignocellulose is the most abundant and inexpensive source of biomass and therefore is widely considered a possible source for liquid fuel. However, processing costs have kept lignocellulosic fuels from competing commercially with starch-based biofuels. In recent years a strategy to protect processing enzymes with synergistic proteins that interact specifically with cellulose has emerged to reduce the amount of enzyme necessary for lignocellulosic biofuel production. Chapter 5 presents the consensus design of a protein based on the family 1 cellulose binding module domain (CBM1) that will bind to cellulosic materials.

1.5 Modifying Human Hair Keratin at the Sequence Level

Biomaterials made from extracted keratin have mechanical and biochemical properties that make them ideal scaffolds for tissue engineering and wound healing. However, naturally extracted keratin contains unwanted byproducts such as melanin. This can be problematic for certain applications of interest, such as ocular tissue regeneration. Recombinant keratin biomaterials are free from these disadvantages. Additionally, heterologous expression of these proteins allows us to manipulate the primary sequence. Chapter six discusses one example of a sequence-level modification to improve the properties of keratin biomaterials through integration of an integrin-binding site.

Chapter 2. Experimental Approach, Techniques, and Instrumentation

2.1 Introduction

Our goal is to create stable protein scaffolds that bind to targets of interest, particularly polysaccharides, as a lack of available targeting agents persists for this class of molecules. This chapter contains a summary of the techniques used to design and characterize these scaffolds.

2.2 Multiple Sequence Alignment and Statistical Design

2.2.1 Development of a Protein Library and Multiple Sequence Alignment

In general, a large and reasonably diverse sequence library of homologous sequences will result in a high-quality Multiple Sequence Alignment (MSA), which in turn should result in a stable protein.[1] Biological databases containing protein sequences searchable by sequence, structure, function, or species enable researchers to build libraries of homologous proteins (**Figure 2.1**).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42
1	L	D	T	K	L	N	I	G	D	C	L	H	D	M	G	K	Y	N	N	A	I	E	V	Y	Y	S	V	D	K	I	L	T	E	T	S	G	I	N	H	P	S	T
2	M	D	T	K	S	N	I	A	F	C	L	N	R	M	E	K	Y	N	E	A	L	G	I	Y	N	S	V	D	K	M	Q	T	E	I	L	G	A	N	H	P	S	T
3	M	H	T	K	I	N	I	A	Y	C	L	S	H	M	G	K	H	N	E	A	L	E	I	Y	Y	S	V	D	K	I	Q	T	Q	K	L	G	I	N	H	P	S	T
...																																										
1285	L	I	A	K	H	N	I	A	S	C	L	Y	N	M	G	K	F	K	E	A	L	E	I	Y	Y	Y	V	E	K	V	K	T	E	N	L	G	A	D	H	P	L	L

Figure 2.1. Example Multiple Sequence Alignment. Sequences were selected from the NCBI database and aligned in Microsoft Excel.

One of the most widely used programs is the Basic Local Alignment Search Tool (BLAST),[2] which calculates the statistical similarity between nucleotide or protein sequences. This sequence search is useful in identifying proteins with a similar sequence to one of interest, or in identifying likely protein domains within a known sequence.

Databases such as the Pfam[3] and Simple Modular Architecture Research Tool (SMART) databases[4] are efficient tools to search structures. They use hidden Markov models (HMMs) built from multiple sequence alignments of known protein domains to detect protein domains in new protein sequences. This allows researchers to input a structure of interest and get back sequences that fit the structure, or a set of rules to apply to larger collections such as the NCBI sequence database.[5]

Tools that identify proteins by function include the Carbohydrate-Active enZYmes (CAzy) database,[6] which contains the functional domains of known enzymes that create, modify, or break glycosidic bonds. We used this database to build a library of Carbohydrate-Binding Module 1 domains in chapter 4.

For MSAs of very small proteins or single repeats such as those discussed in this dissertation rather than full-length proteins, researchers often choose only sequences with no gaps or extra loops.[7-10] This likewise allows us to avoid making decisions about which positions will be stabilizing.

2.2.2 Consensus Design

Consensus design takes the most common amino acid at each position in the multiple sequence alignment. This approach has been successful in improving the stability of a number of proteins.[1, 7-9, 11-21] In general, point mutations of an amino acid residue in a given protein to the consensus residue are stabilizing approximately 50% of the time, neutral about 10% of the time, and destabilizing the remaining 40% of the time.[1] Creating a *de novo* sequence allows researchers to avoid having to predict which mutations will be stabilizing, and has resulted in sequences that have melting temperatures 10–30 °C higher than those in the MSA.[1, 9]

2.2.3 Other Approaches: Global Propensity, Statistical Free Energy

An underlying assumption to consensus protein design is that residue conservation is related to the thermodynamic stability.[10] However, a number of other selective pressures, including the need to express well and avoid aggregation and premature protease digestion, affect how frequently different amino acids occur; proteins likewise only need to be as thermodynamically stable as is required for their function. To attempt to tease out the differences between stabilizing residues and residues that occur due to other factors, researchers can use global propensity ($P_{G,xi}$) (**Equation 2.1**),[22] free energies (ΔG_{xi}) (**Equation 2.2**),[10] or relative entropies (D_{xi}) (**Equation 2.3**)[7] rather than straight consensus. In these equations, kT^* is an arbitrary energy unit, $p(x_i)$ is the probability of amino acid x at position i, and q(x) is the probability of finding that amino acid in a reference database such as the yeast proteome or the Pfam database.

$$P_{G,x_i} = \frac{p(x_i)}{q(x)} \tag{2.1}$$

$$\Delta G_{x_i} = kT^* \sqrt{\sum_x (\ln \frac{p(x_i)}{q(x)})^2} \tag{2.2}$$

$$D_{x_i} = p(x_i) \ln \frac{p(x_i)}{q(x)} \tag{2.3}$$

By relating how often each amino acid occurs at each position in the MSA to how often it appears in a reference database, residues that are truly important for function or thermodynamic stability are selected for over those residues which are selected for other reasons.

2.2.4 Mutual Information Analysis: Correlations Between Positions

Another implicit assumption in consensus design is that each position of the MSA is independent of all other positions.[7] However, packing in the hydrophobic core of the protein

necessarily indicates that each position cannot be independent, as sterics, salt bridge and hydrogen bonding interactions all rely on two or more amino acids (or amino acids of complementary size or type) occurring simultaneously. Correlations between positions can be calculated using mutual information (MI), which is the relative entropy between the actual probability distribution for two positions and the independent probability distribution between those same positions. The MI between two sites i and j is equation 2.4, where $p(x_i)$ is the probability of observing amino acid x at position i, $p(y_j)$ is the probability of observing amino acid y at position j, and $p(x_i, y_j)$ is the probability of observing amino acid x at position i and amino acid y at position j.

$$MI_{ij} = \sum_i \sum_j p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \quad \textbf{(2.4)}$$

To determine the statistical significance of the MI values, we scramble the amino acids in each position and calculate the MI of this scrambled dataset. The highest value from this scrambled data is typically considered the noise. From there, the MI values are broken into groups based on multiples of the standard deviation above the noise.

2.3 Gene Design, Synthesis, and Manipulation

2.3.1 Gene Design and Synthesis

Genes of interest were designed by reverse translation of the protein sequence using the Bioinformatics Sequence Manipulation Suite to codons optimized for *E. Coli* expression.[23] Genes were produced by Klenow extension of overlapping nucleotides purchased from Integrated DNA Technologies (Coralville, IA), and a subsequent polymerase chain reaction (PCR) with two additional amplification primers corresponding to the ends of the gene and

restriction sites for cloning (**Figure 2.2**).[24] Genes were cut at appropriate restriction sites cloned by ligation using T4 DNA ligase into the *pProExHTAm* plasmid, which contains an ampicillin resistance gene, *trc* promoter for gene expression, and N-terminal hexahistidine tag to allow for Ni-NTA purification of the resulting protein.

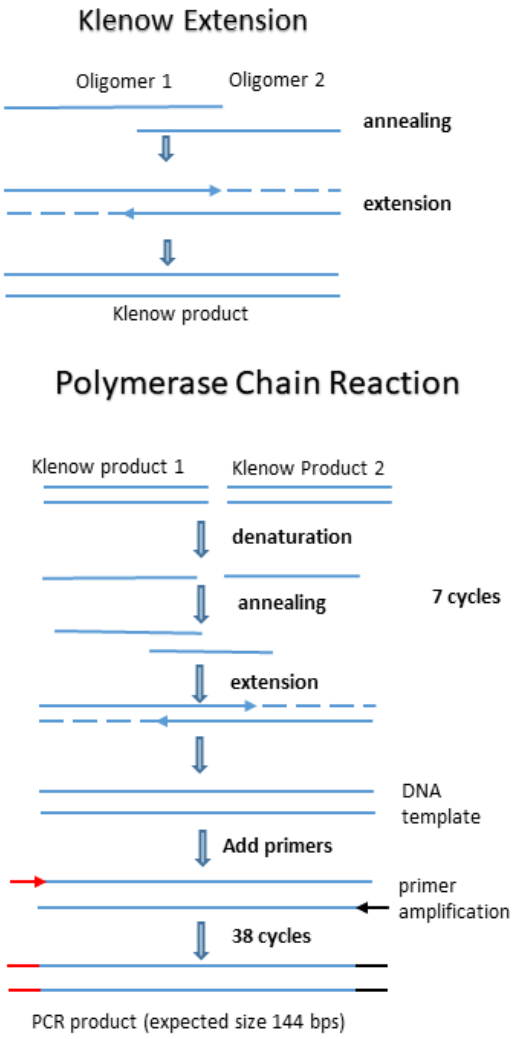


Figure 2.2. Klenow extension and PCR schematic for gene synthesis.

2.3.2 Gene manipulation by Site-Directed Mutagenesis

Modification, insertion, or deletion of a small number of base pairs (approximately 1-3), was accomplished through site-directed mutagenesis (**Figure 2.3**).[25] This PCR technique requires a mismatch primer that overlaps the mutation by at least 9 base pairs upstream and downstream of the mutation, and a high-fidelity polymerase. The mismatch primer anneals to the template vector and gets extended to the full vector length. After approximately 25–35 rounds of PCR, the reaction mixture is digested with DPN1 (New England Biolabs, Ipswich, MA), which digests the naturally synthesized DNA template between methylated adenosine bases in the sequence GATC, leaving only the PCR-synthesized mutagenesis product.

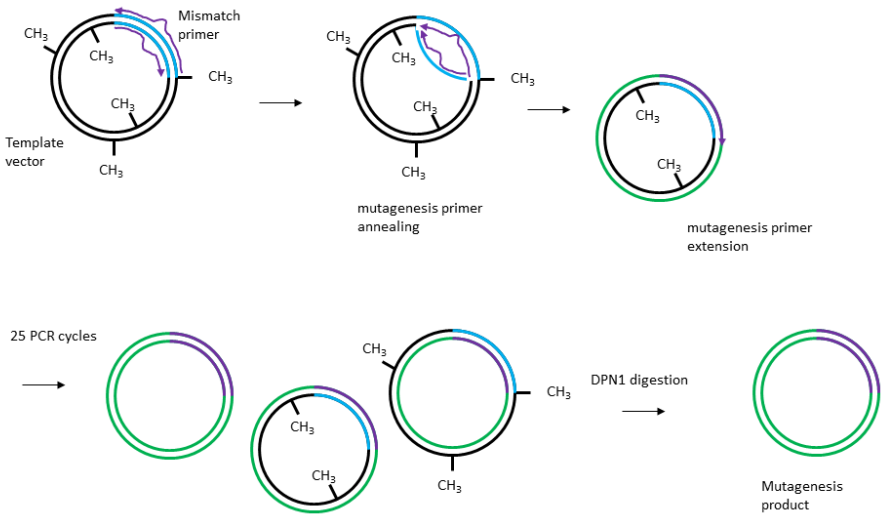


Figure 2.3. Site-directed mutagenesis schematic for changing a small number of base pairs.

2.3.3 Gene manipulation by Megaprimer Whole-Plasmid Cloning

To modify a larger number of base pairs, we relied on the Megaprimer whole-plasmid cloning technique (**Figure 2.4**).[26-27] This two-step PCR technique requires a mutagenic primer with at least 15 base pairs downstream of the reaction and an overlapping primer upstream of the desired mutation. After standard PCR with these primers and template plasmid using a high-fidelity polymerase, a DNA agarose gel with ethidium bromide to confirm synthesis of the

Megaprimer; upon confirmation, the product was cut and purified with a Qiagen gel extraction kit.

The second round of PCR used the Megaprimer product of the first round of PCR along with more template plasmid. As with site-directed mutagenesis above, the template plasmid was then digested with DPN1.

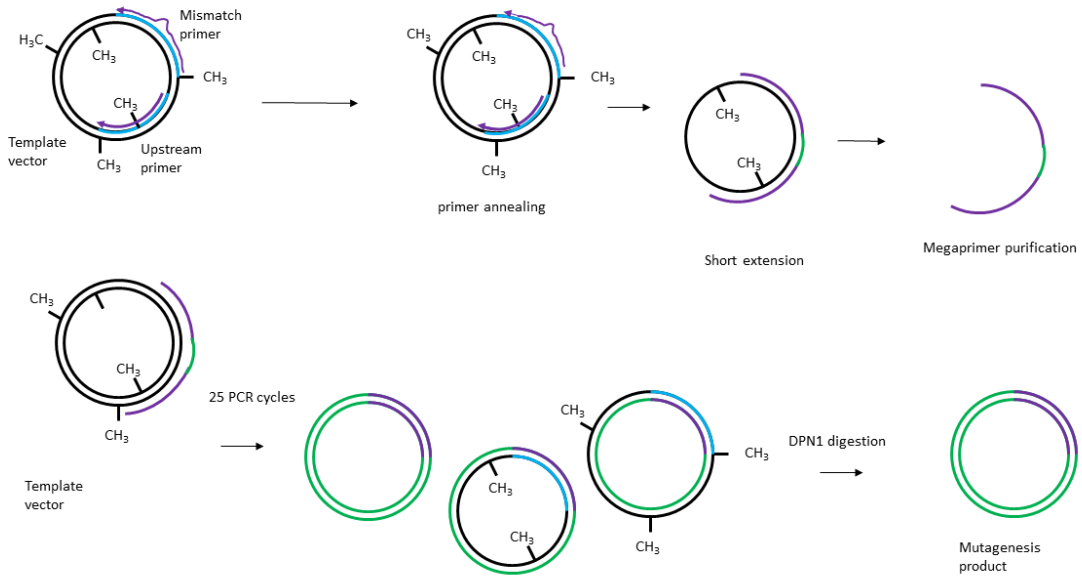


Figure 2.4. Schematic of Megaprimer whole-plasmid cloning technique.

2.4 Protein Expression

After ligation or PCR, the reaction mixture was transformed using electroporation into electrocompetent NEB-10 *E. coli* cells, designed for plasmid preparation. Once a plasmid containing the desired gene was confirmed, it was transformed again into electrocompetent NEB-10 *E. coli* cells to grow more plasmid, and BL-21 *E. coli* cells, designed for protein expression.

The BL-21 cells are grown in Luria Broth media overnight. The overnight culture is then diluted 1:100 in Luria broth, grown to an optical density as measured by UV/Vis at 600 nm of

0.4–0.8, at which point protein expression is induced by isopropyl β-D-1-thiogalactopyranoside (IPTG). Protein expression occurs at 37°C for 3–4 hours, 28°C for 6 hours, or 17°C for 12–24 hours.

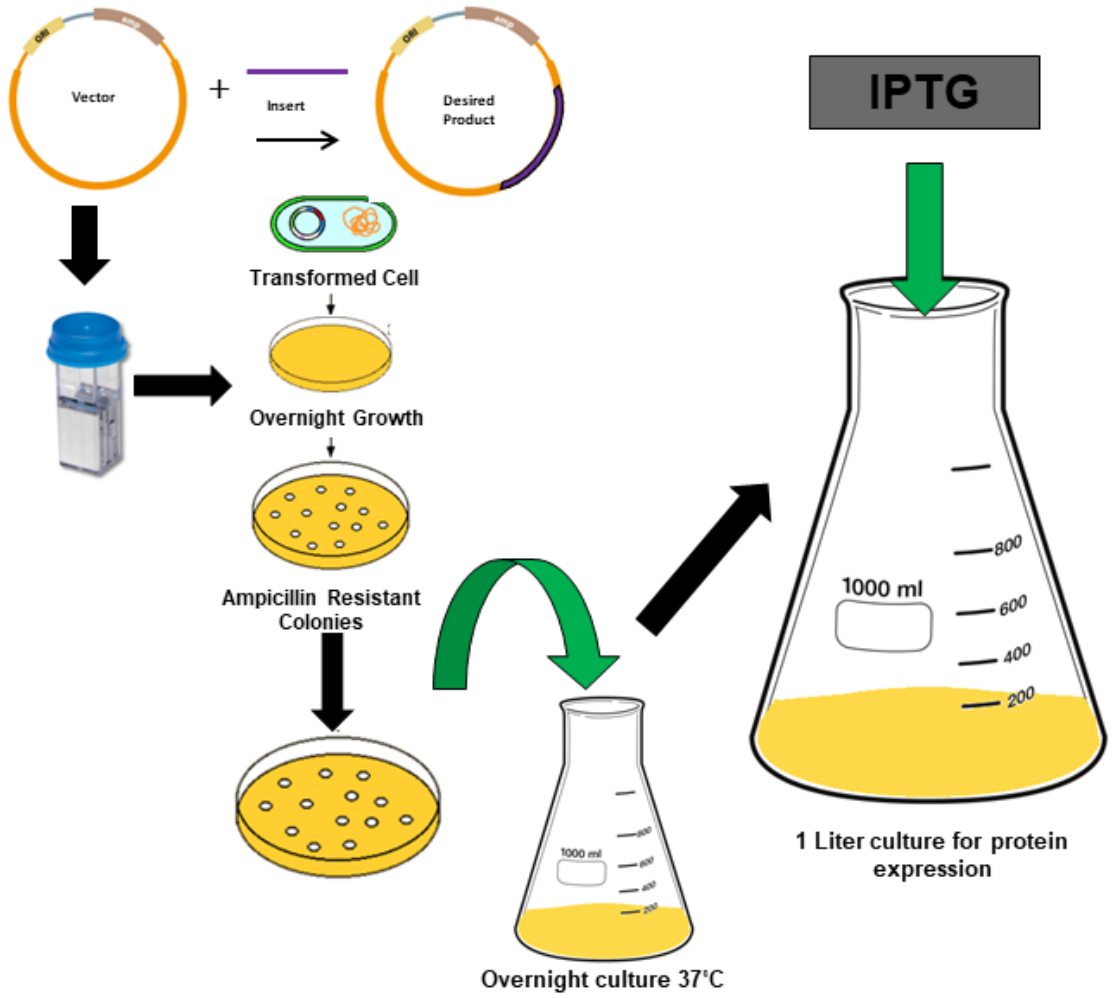


Figure 2.5. Schematic of protein expression.

2.5 Protein Purification by Affinity Chromatography

Proteins are commonly purified through affinity chromatography (**Figure 2.5**).[28] The hexahistidine tag on the N termini of our protein binds to nickel with a higher affinity than histidines in *E. coli* proteins due to avidity effects, and therefore will remain bound to the resin as more loosely bound proteins are washed off the resin. Ni-NTA can be performed under native

or denaturing conditions and tolerates small amounts of reducing agents such as β-mercaptoethanol.

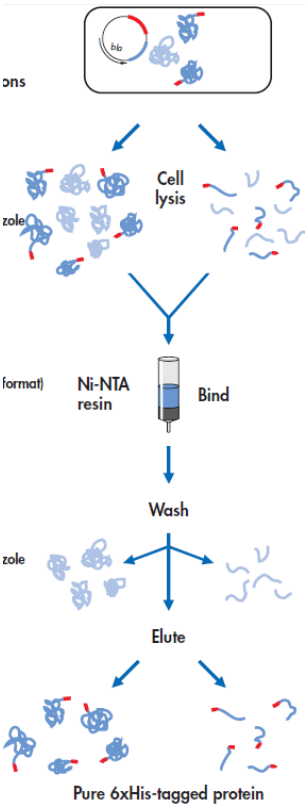


Figure 2.6. Ni-NTA purification of hexahistidine-tagged protein

2.6 Protein Characterization

2.6.1 Size-Exclusion Chromatography (SEC)

SEC allows us to estimate protein size, which in turn will give us information about the folding and oligomeric state of the protein. Size-exclusion columns used for protein purification are packed with porous agarose beads. These beads will trap smaller molecules in their pores while larger molecules will move through the void volume between the beads, allowing separation of molecules based on size. To estimate the size of our protein, we created a calibration curve from known protein standards, which will relate the logarithm of the molecular weight to the ratio of eluted volume/void volume.

2.6.2 Secondary Structure Characterization

We used far-UV Circular Dichroism (CD) spectroscopy to characterize the secondary structure of our proteins. CD generally refers to the unequal absorption of circularly polarized light by chiral chromophores. The amide bonds of protein backbones have characteristic CD spectra depending on their fold. α-Helical proteins have characteristic negative bands at 220 and 208 nm, while β-sheets have a characteristic negative band at 218 nm. Unfolded or random coil proteins meanwhile have very little signal above 210 nm, and characteristic negative bands around 195 nm.[29] This data can be used to estimate the structure of novel proteins.

2.6.3 Measuring Thermal Stability

Circular dichroism can also be used to measure the thermal stability of a protein. When the CD spectra of a protein is known, an appropriate peak characteristic of the secondary structure (such as 220 nm for an α-helical protein) can be chosen and monitored over a given temperature range. The extent of unfolding will be given by **Equation 2.5**, where α is the fraction unfolded at any temperature, θ_T is the observed ellipticity at temperature T, θ_F is the ellipticity of the folded form of the protein and θ_U is the ellipticity of the fully unfolded form of the protein at the appropriate wavelength.[30]

$$\alpha = (\theta_F - \theta_T) / (\theta_F - \theta_U) \quad \textbf{(2.5)}$$

2.7 Binding Characterization

Binding characterization can be accomplished using a number of different techniques, depending on the specifics of the protein and its ligand, including fluorescence anisotropy, isothermal titration calorimetry, and surface plasmon resonance. In our work, we used a UV/Vis

technique that indirectly measured the binding affinity to cellulose, and a Quartz Crystal Microbalance to measure the adsorption onto a regenerated cellulose thin film.

2.7.1 UV/Vis to Characterize Binding to a Cellulose Suspension

UV/Vis was used to characterize binding of our cellulose binding module protein to a cellulose suspension. Protein concentration was measured by UV/vis at 280 nm using extinction coefficients calculated from the protein sequence by the Expasy Protparam tool.[31] Protein samples of known concentration were mixed with an aqueous suspension of Avicel microcrystalline cellulose by rotation at 4°C for 20 hours. The cellulose was sedimented by centrifugation, along with bound protein. The supernatant with unbound protein was collected and protein concentration was again measured at 280 nm. The difference between the starting protein concentration and the protein concentration after mixing was taken to be the amount of protein bound to the cellulose. This allowed us to indirectly measure the binding affinity of our protein to the Avicel.

2.7.2 Quartz Crystal Microbalance with Dissipation Monitoring (QCM-D) to Measure Binding Affinity

We used QCM-D to measure the molecular adsorption of protein onto a regenerated cellulose surface. QCM-D directly measures the frequency of a quartz crystal, which can be related to adsorbed mass by the Sauerbrey equation (**Equation 2.6**), where C is a constant and n is the overtone number. In this equation, Δm is the change in mass per unit area on the sensor corresponding to a measured change in frequency (Δf).

$$\Delta m = -\frac{C\Delta f}{n} \qquad \textbf{(2.6)}$$

We measured the mass of protein absorbed onto a cellulose thin film by flowing our protein into the QCM-D cell until equilibrium, then rinsing off reversibly bound protein. This allowed us to directly measure the binding of our protein to the regenerated cellulose thin film.

2.8 References

1. Porebski, B. T.; Buckle, A. M., Consensus protein design. *Protein Eng Des Sel* **2016,** *29* (7), 245-251.
2. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J., Basic local alignment search tool. *J Mol Biol* **1990,** *215* (3), 403–10.
3. Sonnhammer, E. L.; Eddy, S. R.; Birney, E.; Bateman, A.; Durbin, R., Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* **1998,** *26* (1), 320-2.
4. Schultz, J.; Milpetz, F.; Bork, P.; Ponting, C. P., SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A* **1998,** *95* (11), 5857–64.
5. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **2014,** *42* (Database issue), D7-17.
6. The Carbohydrate-Active enZYmes Database. : ; http://www.cazy.org.
7. Durani, V.; Magliery, T. J., Protein engineering and stabilization from sequence statistics: variation and covariation analysis. *Methods Enzymol* **2013,** *523*, 237–56.
8. Sullivan, B. J.; Nguyen, T.; Durani, V.; Mathur, D.; Rojas, S.; Thomas, M.; Syu, T.; Magliery, T. J., Stabilizing Proteins from Sequence Statistics: The Interplay of Conservation and Correlation in Triosephosphate Isomerase Stability. *J Mol Biol* **2012,** *420* (4–5), 384–399.
9. Sullivan, B. J.; Durani, V.; Magliery, T. J., Triosephosphate Isomerase by Consensus Design: Dramatic Differences in Physical Properties and Activity of Related Variants. *J Mol Biol* **2011,** *413* (1), 195-208.
10. Magliery, T. J.; Regan, L., Beyond consensus: statistical free energies reveal hidden interactions in the design of a TPR motif. *J Mol Biol* **2004,** *343* (3), 731-45.
11. Cunha, E. S.; Hatem, C. L.; Barrick, D., Synergistic enhancement of cellulase pairs linked by consensus ankyrin repeats: Determination of the roles of spacing, orientation, and enzyme identity. *Proteins* **2016,** *84* (8), 1043-1054.
12. Parker, R.; Mercedes-Camacho, A.; Grove, T. Z., Consensus design of a NOD receptor leucine rich repeat domain with binding affinity for a muramyl dipeptide, a bacterial cell wall fragment. *Protein Sci* **2014,** *23* (6), 790–800.
13. Tiede, C.; Tang, A. A. S.; Deacon, S. E.; Mandal, U.; Nettleship, J. E.; Owen, R. L.; George, S. E.; Harrison, D. J.; Owens, R. J.; Tomlinson, D. C.; McPherson, M. J., Adhiron: a stable and versatile peptide display scaffold for molecular recognition applications. *Protein Engineering, Design and Selection* **2014,** *27* (5), 145-155.
14. Kunik, V.; Peters, B.; Ofran, Y., Structural Consensus among Antibodies Defines the Antigen Binding Site. *PLoS Comput Biol* **2012,** *8* (2), e1002388.
15. Lee, S.-C.; Park, K.; Han, J.; Lee, J.-j.; Kim, H. J.; Hong, S.; Heu, W.; Kim, Y. J.; Ha, J.-S.; Lee, S.-G.; Cheong, H.-K.; Jeon, Y. H.; Kim, D.; Kim, H.-S., Design of a binding scaffold based on variable lymphocyte receptors of jawless vertebrates by module engineering. *Proceedings of the National Academy of Sciences* **2012,** *109* (9), 3299.
16. Wezner-Ptasińska, M.; Krowarsch, D.; Otlewski, J., Design and characteristics of a stable protein scaffold for specific binding based on variable lymphocyte receptor sequences. *BBA-Proteins Proteom* **2011,** *1814* (9), 1140-1145.

17. Kajander, T.; Cortajarena, A. L.; Regan, L., Consensus design as a tool for engineering repeat proteins. *Methods Mol Biol* **2006,** *340*, 151–70.
18. Forrer, P.; Binz, H. K.; Stumpp, M. T.; Pluckthun, A., Consensus design of repeat proteins. *Chembiochem* **2004,** *5* (2), 183–9.
19. Binz, H. K.; Stumpp, M. T.; Forrer, P.; Amstutz, P.; Pluckthun, A., Designing repeat proteins: well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *J Mol Biol* **2003,** *332* (2), 489–503.
20. Lehmann, M.; Wyss, M., Engineering proteins for thermostability: the use of sequence alignments versus rational design and directed evolution. *Curr Opin Biotechnol* **2001,** *12* (4), 371-5.
21. Knappik, A.; Ge, L.; Honegger, A.; Pack, P.; Fischer, M.; Wellnhofer, G.; Hoess, A.; Wolle, J.; Pluckthun, A.; Virnekas, B., Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. *J Mol Biol* **2000,** *296* (1), 57–86.
22. Main, E. R.; Xiong, Y.; Cocco, M. J.; D'Andrea, L.; Regan, L., Design of stable alpha-helical arrays from an idealized TPR motif. *Structure* **2003,** *11* (5), 497–508.
23. Stothard, P., The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques* **2000,** *28* (6), 1102, 1104.
24. Holowachuk, E. W.; Ruhoff, M. S., Efficient gene synthesis by Klenow assembly/extension-Pfu polymerase amplification (KAPPA) of overlapping oligonucleotides. *PCR methods and applications* **1995,** *4* (5), 299-302.
25. Ho, S. N.; Hunt, H. D.; Horton, R. M.; Pullen, J. K.; Pease, L. R., Site-directed mutagenesis by overlap extension using the polymerase chain reaction. *Gene* **1989,** *77* (1), 51-59.
26. Tyagi, R.; Lai, R.; Duggleby, R. G., A new approach to 'megaprimer' polymerase chain reaction mutagenesis without an intermediate gel purification step. *BMC biotechnology* **2004,** *4*, 2.
27. Vander Kooi, C. W., Chapter Twenty One - Megaprimer Method for Mutagenesis of DNA. In *Methods Enzymol*, Lorsch, J., Ed. Academic Press: 2013; Vol. 529, pp 259-269.
28. Spriestersbach, A.; Kubicek, J.; Schafer, F.; Block, H.; Maertens, B., Purification of His-Tagged Proteins. *Methods Enzymol* **2015,** *559*, 1-15.
29. Greenfield, N. J., Using circular dichroism spectra to estimate protein secondary structure. *Nat Protoc* **2006,** *1* (6), 2876-2890.
30. Greenfield, N. J., Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions. *Nat Protoc* **2006,** *1* (6), 2527-2535.
31. Gasteiger, E.; Hoogland, C.; Gattiker, A.; Duvaud, S. e.; Wilkins, M. R.; Appel, R. D.; Bairoch, A., Protein Identification and Analysis Tools on the ExPASy Server. In *The Proteomics Protocols Handbook*, Walker, J. M., Ed. Humana Press: Totowa, NJ, 2005; pp 571-607.

Chapter 3: Engineering Non-Immunoglobulin Proteins of the Innate Immune System

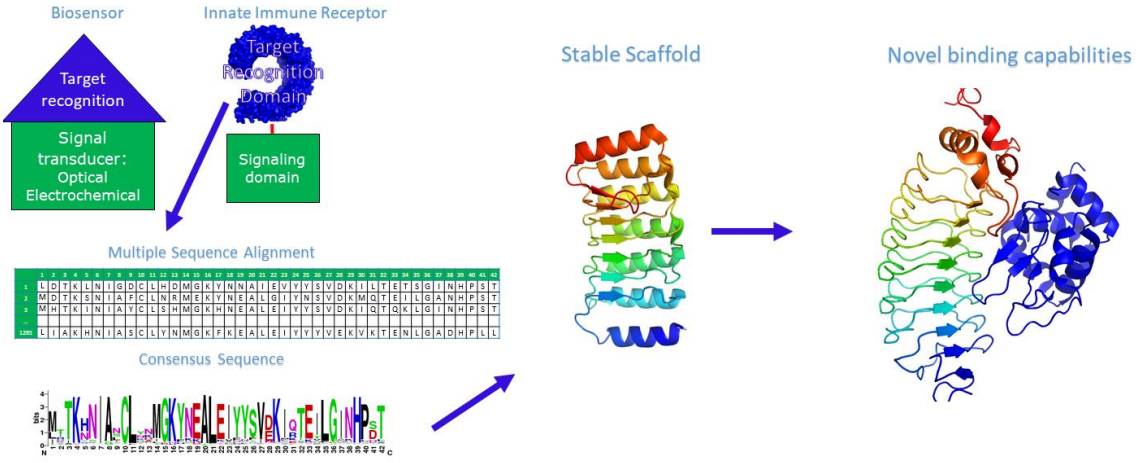
(Submitted)

Jennifer P. McCord and Tijana Z. Grove

Department of Chemistry, Virginia Tech, Blacksburg, VA 24061

Keywords: repeat proteins, leucine-rich repeats, NOD-like receptors, variable lymphocyte receptors, protein engineering, engineered binding scaffolds

3.1 Abstract



Problems with stability, expense of expression, and specificity of interaction of Ig antibodies have led to the development of alternative binding scaffolds. Repeat proteins have emerged as one promising class of scaffolds, but often are limited to protein and peptide targets. In recent years, researchers have looked to non-immunoglobulin proteins of the immune system for inspiration for binding scaffolds that can bind glycans and other classes of biomolecules in addition to proteins. These proteins are stable and monomeric, with elongated and highly

variable binding surfaces. Their ability to target glycans and glycoproteins fill an important gap in current tools for research and biomedical applications.

3.2 Introduction

Many applications in biomedicine and research require recognition agents with specificity and selectivity for their target. Immunoglobulin (Ig) antibodies are the best-studied examples of such recognition agents, and have been reviewed extensively.[1-3] However, problems with stability, expense of expression, and specificity of interaction of Ig antibodies have led in recent years to the development of alternative binding scaffolds.[4-8] Antibodies rely on disulfide bond formation for their stability, requiring difficult, expensive eukaryotic cell production; this also makes them unsuitable for applications in reducing conditions.[9] The binding sites on antibodies are such a small part of a large protein that nonspecific interactions occur with some frequency. A 2008 study found that among 6,000 routinely used antibodies, fewer than half were specific binders for their intended target.[10-11]

In the search for better recognition agents, repeat proteins have emerged as one promising class of scaffolds. Designed repeat proteins are comparatively simple to design and can have highly variable binding surfaces, and several examples of designed repeat proteins are remarkably stable.[12-18] Repeat proteins can be expressed in *E. coli,* can be engineered to be well-folded and cysteine-free, and have very large binding sites relative to the size of the protein, minimizing available surface for nonspecific binding.[6] Additionally, their modular architecture makes them relatively simple to analyze and redesign.[13] Designed repeat proteins are therefore well-suited for a wide variety of biomedical applications. DARPins (Designed Ankyrin Repeat Proteins) in particular have shown remarkable promise for in-situ diagnostics,[19-20] and two DARPin drugs have reached clinical trials.[21-22] However, a drawback of DARPins and other

commonly utilized repeat protein scaffolds is that their targets are typically restricted to proteins or peptides,[23] limiting available targets of interest.

Interestingly, nature uses repeat proteins in its biosensors in the innate immune system throughout the biosphere,[24] as well as the adaptive immune system in jawless vertebrates.[25] The repeat domains in the innate immune system recognize as diverse biomolecules as glycans and nucleic acids as well as proteins characteristic of pathogens and cellular damage.[26] These repeat proteins are therefore a promising starting point for designing repeat proteins that can bind to diverse biomolecular targets.

Repeat proteins are also found in the adaptive immune system proteins of lampreys and hagfish, the last surviving jawless vertebrates. Their unusual adaptive immune system functions similarly to other vertebrates, except that its scaffolds for pathogen defense are repeat proteins known as variable lymphocyte receptors (VLRs) rather than immunoglobulins.[27] The modular nature of repeat proteins allows for repeats to be shuffled, inserted, and deleted, and variable lymphocyte receptor DNA takes advantage of this modularity.[28] Scaffolds based on VLRs have been utilized by researchers to recognize glycan[29-31] as well as protein targets.[32-35]

This review will focus on the design of scaffolds based on non-immunoglobulin proteins of the immune system that use repeat domains to recognize their targets. We will begin with the brief general summary of engineered repeat proteins and then discuss recent engineered immune system proteins with an emphasis on the specific advantages of these scaffolds afforded by the modularity of repeat proteins.

3.3 Engineering Repeat Proteins

Repeat proteins are the most abundant natural binding proteins after immunoglobulins, likely as a result of their versatility and binding affinities.[36] The modularity of repeat proteins allows for insertions, deletions, and shuffling of individual repeats to create novel binding capabilities while maintaining the overall protein structure.[17, 28] Further, they are comparatively easy to design, have highly variable binding surfaces, and can be remarkably stable.[12-18]. Engineered binding scaffolds based on repeat proteins have found applications in biosensing,[37-39] diagnostics,[37, 40] and therapeutics.[5, 41-46]

Repeat domain architecture is characterized by tandem arrays of small structural motifs, typically 20–40 amino acids in length.[47-49] Repeat domains are versatile, stable binding domains reliant only on short-range interactions for their stability. This architecture results in an extended structure with a large surface area and an elongated binding interface. Many proteins containing repeat domains mediate protein-protein interactions, however, we find immune system repeat proteins particularly interesting due to their ability to recognize nonprotein targets.

Successful binding repeat proteins have been designed using methods that run the gamut from library screening[20] to grafting a known binding site onto a new scaffold.[45, 50] One oft-employed strategy to obtain a stable scaffold is consensus design, which takes the most common amino acid from each position in the protein sequence from a known library of proteins.[51] With repeat proteins, consensus design is often done for just one repeat that is then repeated as many times as desired for a stable scaffold, or one that will bind the target of interest most tightly, with modifications often made for N and C terminal (or "capping") repeats. This method increases the stability of the scaffold, which in turn can allow for more stable libraries of binding variants.[16, 18]

3.4 Non-Immunoglobulin Immune System Proteins

3.4.1 The Innate Immune System

Organisms use the system of proteins that comprise the immune system to detect and respond to pathogenic attack. Receptors of the innate immune system non-specifically but immediately recognize molecules indicative of pathogenic attack or other threats. Plants and mammals use similar domain combinations among their innate immune proteins and recent work suggests a similar system of proteins in fungi as well.[52] Based on phylogenetic analysis data, these proteins likely evolved through convergent, rather than divergent evolution.[53] Startling similarities among these proteins throughout the biosphere are a testament to their effectiveness at host defense, among their other functions.[54] While these parallels suggest the possibility of a common ancestor, recent research indicates that this is the result of convergent rather than divergent evolution.[55] Interestingly, the ligand recognition domain of these proteins is a repeat-protein motif. As nature's biosensors, innate immune system proteins are excellent inspiration for molecular recognition scaffolds.

3.4.1.1 Toll-like Receptors and NOD-like Receptors

Animals have two major classes of innate immune receptor: Toll-like receptors (TLRs) and NOD-like receptors (NLRs). These respectively correspond to extracellular and intracellular defense, each using leucine-rich repeats to recognize their targets.

3.4.1.1.1 Toll-Like Receptors

TLRs, a family of membrane-bound proteins, form the basis for extracellular pattern recognition in vertebrates. TLRs are the first discovered and best-studied family of innate immune proteins. Their structure is composed of an extracellular leucine-rich repeat (LRR)

domain and an intracellular Toll-interleukin receptor (TIR) domain. The LRR domain is the biorecognition agent, while the TIR domain initiates the immune response within the cell.

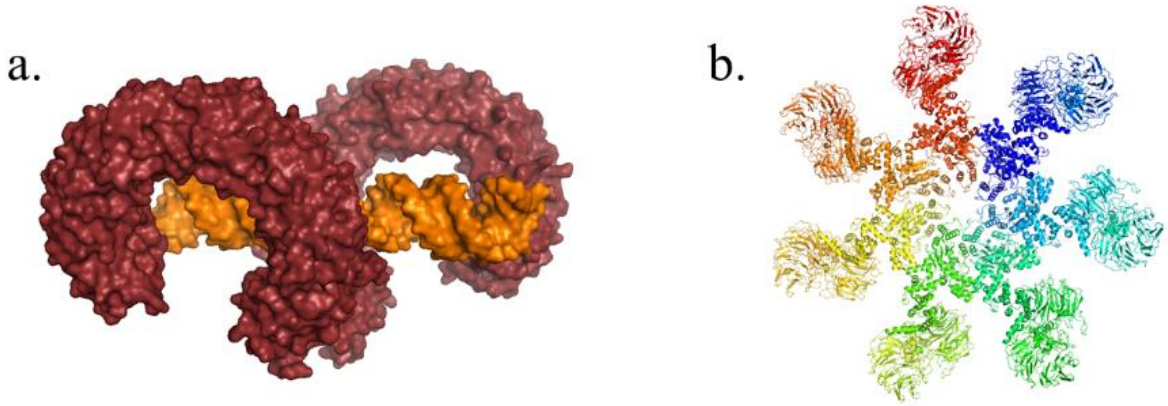


Figure 3.1. a. Example of the dimeric binding mode of TLRs: mouse TLR3 ectodomain in dimeric complex with its ligand complexed with double-stranded RNA. PDB: 3CIY b. A model of how NLRs oligomerize, using Apaf-1, a homologous protein.[56] This shows oligomerization of the non-binding domains into a heptamer, leaving ligand recognition domains free for individual monomeric binding.

TLRs recognize a wide variety of ligands, including varieties of DNA, RNA, lipoproteins, glycolipids, proteins, and other toxins.[57-64] However, TLRs always rely on a dimeric mode of binding; that is, two proteins are required to bind one ligand (**Figure 3.1a**). Most TLRs form homodimers to bind their ligands, but mammalian TLR2 must form heterodimers with TLR1 or TLR6 in order to recognize certain ligands. In spite of their diverse ligand binding abilities, need for oligomerization in order to bind ligands makes them inconvenient targets for protein design.

3.4.1.1.2 NOD-Like Receptors (NLRs)

NLRs are a family of intracellular proteins that recognize damage-associated molecular patterns (DAMPs) and pathogen-associated molecular patterns (PAMPs) within the cell.[24] Characteristic NLR structure is tripartite, with a C-terminal leucine-rich repeat domain (LRR), a central nucleotide-binding and oligomerization domain (NOD), and an N-terminal effector-binding domain. The N-terminal domain determines the NLR family (Table 1).[65] NLRs use LRRs as binding domains like TLRs do, but NLR LRRs do not dimerize to recognize ligands.[66] While Proell, et al. hypothesize that NLRs likely oligomerize to form a signaling platform similar to Apaf-1, a homologous protein (**Figure 3.1b**),[56] this oligomerization does not affect the mode of binding of the LRR domain with its ligand.

NLR proteins play a particularly important role in cells where TLRs are expressed in low levels or not at all, particularly in the mammalian colon where a large population of symbiotic microbes exist.[67] If TLRs were present in these microbe-rich environments, they would overwhelmingly recognize helpful rather than dangerous microbes. However, microbes that have breached the cell wall pose a clear danger to the cell. In these cells, NLRs are the only defense against pathogenic attacks.

Table 3.1. NLR families and their associated ligands

NLR family	Ligands
NOD (CARD-NOD-LRR Proteins)	Pathogen-associated molecular patterns,[68] including: Peptidoglycan fragments, Muramyl dipeptide (MDP)
NALP (PYD-NOD-LRR Proteins)	Damage-associated molecular patterns,[68] including: Double-stranded RNA from bacteria and viruses, toxins, MDP, uric acid (from apoptotic cells)
NAIP (BIR-NOD-LRR Proteins)	Flagellin, Type III secretion system (TTSS)[69]

Mammalian NLR proteins recognize a wide variety of ligands associated with microbes and cellular damage, summarized in **Table 3.1**. These proteins use the same structural motifs to bind diverse molecules such as peptidoglycans, RNA, and uric acid.

3.4.1.1.3 Novel Domain Combinations in NOD-like Receptors Throughout the Biosphere

Recent reports indicate that some early diverging organisms have large and complex NLR repertoires (**Figure 3.2**).[70-71] However, unlike canonical NLR proteins that must contain both a NOD and an LRR domain, receptors in these organisms have similar domain structures, but with a C-terminal repeat domain other than an LRR. *Hydra magnipapillata*, a freshwater polyp, has a particularly large NOD-like receptor system, containing 121 NLR genes (compared to 22 human NLR genes). Additional research has found that the NBD domains in its NLR genes are phylogenetically related to those in human NLRs.[72]

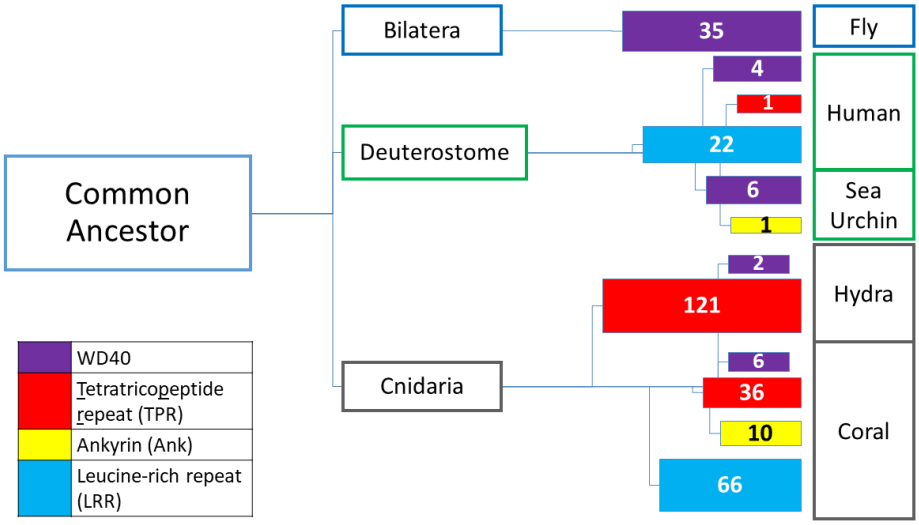


Figure 3.2. Relative number of nucleotide-binding domain- (defined as NBD, NACHT, and NB-ARC) and repeat domain-containing genes in selected organisms throughout the biosphere. Size of the colored box represents the relative number of genes containing each repeat for each organism. Figure adapted from Hamada et al.[70]

This research implies that the majority of NLR proteins in *Hydra* have C-terminal TPRs instead of LRRs. Flies and coral also have a significant number of non-LRR-containing NLR genes. Interestingly, representative early divergent species that do not use LRRs in their NLRs have several times the number of such proteins as humans and other mammals do. We hypothesize that while mammal LRR-containing NLRs can recognize a variety of ligands, other repeat-domain containing NLRs may be specific to one ligand, or class of ligands. These novel domain combinations have the potential to increase the design space for new alternative protein scaffolds able to recognize non-protein targets.

3.4.2 Variable Lymphocyte Receptors (VLRs)

Lampreys and hagfish, the surviving jawless vertebrates, have adaptive immune system proteins that share characteristics of the adaptive and innate immune systems in jawed

vertebrates, as they are highly variable and specific, but use leucine-rich repeat domains as recognition elements.[25] Three distinct types of VLR (A, B, and C) have been identified, with VLRB proteins most commonly used in VLR applications as they are expressed solubly.[73] Alder, et al. estimated a repertoire of 10^{14}–10^{17} unique receptors, comparable to the repertoire of mammalian antibodies.[62]

VLRs have a structure composed of an N-terminal capping LRR (LRRNT), an LRR fragment (LRR1), up to 7 variable LRRs (LRRVs), a terminal LRR (LRRVe), a connecting peptide (CP), a C-terminal capping LRR (LRRCT), and a stalk region rich in threonine and proline that connects the protein to a glycosylphosphatidylinositol (GPI) anchor and cysteine-rich hydrophobic tail. The hydrophobic tail can drive multimerization, which some research has suggested serves to improve the binding specificity,[62, 74] though often, researchers produce monomeric VLRs by expressing only from LRRNT–LRRCT. VLRs are particularly useful tools for researchers for their ability to recognize glycans,[29] which are particularly lacking in biochemical tools for detection.[66]

VLRs are rapidly becoming a useful antibody alternative, and were recently the subject of an excellent comprehensive review by Waters and Shusta.[73] They are able to bind to both proteins and glycans with high specificity and affinity.

3.5 Engineered Binding Scaffolds

3.5.1 Engineered Binding Scaffolds Based on the Innate Immune System

Our lab hypothesized that we could take advantage of NLR proteins' ability to bind medically relevant ligands, and use consensus design to create a stable protein scaffold that would bind non-protein targets of interest. To that end, Parker, et al. used consensus design to

create a protein scaffold based on the leucine-rich repeat domain of NLRs.[66] We created a multiple sequence alignment by searching confirmed human NLR protein sequences from the HUGO database in the NCBI database to find mammalian homologs for a total of 311 individual LRR sequences, further separating these into capping and internal repeats to increase folding and stability. Taking the most common amino acid at each position resulted in a scaffold that is monomeric, stable, and cysteine-free. Without additional affinity maturation, CLRR2 binds stereoselectively with micromolar affinity to muramyl dipeptide.

Marold, et al. recently used consensus design to create a protein based on the repeat domains in NLR-like proteins in the fungus *Podospora anserina*, which contain a partial N-terminal NB-ARC domain with a C-terminal repeat domain of TPRs from family 10, which they have termed 42PRs due to their unusual length of 42 amino acids instead of the typical 34 found in TPRs.[75] They identified a unique 42PR sequence in the fungus *P. anserina*, which consisted of 15 sequences. Their protein consists of a six-repeat consensus of these sequences, with capping helices altered to be more hydrophilic in order to express the proteins solubly. Because of the large number of 42PRs and the high similarity between sequences, the researchers hypothesize a possibility this protein could bind repetitive targets, such as pathogenic DNA.

3.5.2 Engineered Binding Scaffolds based on Variable Lymphocyte Receptors

Monoclonal VLRs have been obtained, most notably by the Cooper, Pancer, and Mariuzza labs, by immunizing lampreys (and more recently, hagfish[76]) to antigens of interest. The cDNA from this immunization can be harvested, and better binding achieved through random mutagenesis and selection through yeast display.[33] Due to the evolutionary distance between jawless vertebrates and mammals, monoclonal VLRs can used in this manner to recognize conserved mammalian glycans that might be difficult for traditional immunoglobulin-

based immunization techniques due to tolerance mechanisms, such as blood group carbohydrates.[29] Libraries generated from VLRB cDNA have also been used to select new binders using surface display platforms.[32] Researchers have successfully produced VLRs using these methods to target proteins,[32, 77] glycoproteins,[76, 78] and glycans,[29-31] as well as whole cells.[34]

To increase the stability and recombinant expression yield of the VLR scaffold, two groups have created consensus VLR scaffolds, dVLRs[79] and repebodies,[80] that can be expressed in *E. coli* and have been used to target proteins involved in a variety of diseases.[35, 81-82] These scaffolds are highly stable, and have shown general applicability to detect a target of interest by high-throughput screening.[32, 83-84]

Wezner-Ptasinska and Otlewski created the dVLR scaffold using consensus design of an entire VLRB protein, defined as containing one LRRV module, by aligning 222 unique sequences from the SMART and UniProt databases in ClustalW.[79] Antigen recognition positions were considered to be those with the highest variability, as well as those in close contact with an antigen in two crystal structures. Using these positions, they created a library and used phage display to select for variants binding to hen egg white lysozyme and psoriasin as model proteins.[83]

Lee, et al. also used consensus design for their repebodies, but limited their consensus to one LRRV module, from 1439 LRR modules from the UniProt and NCBI databases. The researchers used the natural C-terminal capping motifs, substituting the natural N-terminal cap of a VLR with the N-terminal cap from internalin B to increase *E. coli* expression, and varied the number of LRRV modules from 3-6.[80] Repebody binding reagents were first created through rational design by selecting and changing residues responsible for the interactions with known VLR-antigen pairs— myeloid differentiation protein-2 and hen egg white lysozyme, resulting in

nano- and micromolar binding affinities, respectively. Researchers subsequently used phage display to select repebodies to bind other proteins, with binding affinities in the nanomolar range,[35, 85] and were able to use rational affinity maturation to improve binding into the picomolar range.[82]

3.6 Summary

We have presented an overview of protein engineering efforts based non-immunoglobulin proteins of the immune system. Based on repeat protein domains, these proteins are simple to design and modular, with highly variable binding surfaces. Compared with immunoglobulins, these proteins offer high affinity and often superior specificity for their targets. While many protein engineering efforts are only able to target proteins and peptides, these engineered immune proteins are able to target glycans as well, filling an important gap in tools for research and biomedical applications.

3.7 References

1. Lambert, J. M.; Morris, C. Q., Antibody–Drug Conjugates (ADCs) for Personalized Treatment of Solid Tumors: A Review. *Advances in Therapy* **2017,** *34* (5), 1015-1035.
2. Bordeaux, J.; Welsh, A.; Agarwal, S.; Killiam, E.; Baquero, M.; Hanna, J.; Anagnostou, V.; Rimm, D., Antibody validation. *BioTechniques* **2010,** *48* (3), 197-209.
3. Carter, P. J.; Lazar, G. A., Next generation antibody drugs: pursuit of the 'high-hanging fruit&#39. *Nature Reviews Drug Discovery* **2017,** *17*, 197.
4. Könning, D.; Kolmar, H., Beyond antibody engineering: directed evolution of alternative binding scaffolds and enzymes using yeast surface display. *Microbial Cell Factories* **2018,** *17* (1), 32.
5. Gebauer, M.; Skerra, A., Alternative Protein Scaffolds as Novel Biotherapeutics. In *Biobetters: Protein Engineering to Approach the Curative*, Rosenberg, A.; Demeule, B., Eds. Springer New York: New York, NY, 2015; pp 221-268.
6. Fiedler, M.; Skerra, A., Non-Antibody Scaffolds as Alternative Therapeutic Agents. In *Handbook of Therapeutic Antibodies*, Wiley-VCH Verlag GmbH & Co. KGaA: 2014; pp 435-474.
7. Boersma, Y. L.; Pluckthun, A., DARPins and other repeat protein scaffolds: advances in engineering and applications. *Curr Opin Biotechnol* **2011,** *22* (6), 849-57.
8. Grönwall, C.; Ståhl, S., Engineered affinity proteins—Generation and applications. *J Biotechnol* **2009,** *140* (3–4), 254–269.
9. Nygren, P. A.; Skerra, A., Binding proteins from alternative scaffolds. *J Immunol Methods* **2004,** *290* (1–2), 3–28.
10. Berglund, L.; Björling, E.; Oksvold, P.; Fagerberg, L.; Asplund, A.; Al-Khalili Szigyarto, C.; Persson, A.; Ottosson, J.; Wernérus, H.; Nilsson, P.; Lundberg, E.; Sivertsson, Å.; Navani, S.; Wester, K.; Kampf, C.; Hober, S.; Pontén, F.; Uhlén, M., A Genecentric Human Protein Atlas for Expression Profiles Based on Antibodies. *Mol Cell Proteomics* **2008,** *7* (10), 2019–2027.
11. Bradbury, A.; Pluckthun, A., Reproducibility: Standardize antibodies used in research. *Nature* **2015,** *518* (7537), 27-9.
12. Sawyer, N.; Chen, J.; Regan, L., All repeats are not equal: a module-based approach to guide repeat protein design. *J Mol Biol* **2013,** *425* (10), 1826-38.
13. Javadi, Y.; Itzhaki, L. S., Tandem-repeat proteins: regularity plus modularity equals design-ability. *Curr Opin Struct Biol* **2013,** *23* (4), 622–631.
14. Kajava, A. V., Tandem repeats in proteins: from sequence to structure. *J Struct Biol* **2012,** *179* (3), 279-288.
15. Main, E. R.; Xiong, Y.; Cocco, M. J.; D'Andrea, L.; Regan, L., Design of stable alpha-helical arrays from an idealized TPR motif. *Structure* **2003,** *11* (5), 497–508.
16. Binz, H. K.; Stumpp, M. T.; Forrer, P.; Amstutz, P.; Pluckthun, A., Designing repeat proteins: well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *J Mol Biol* **2003,** *332* (2), 489–503.
17. Forrer, P.; Stumpp, M. T.; Binz, H. K.; Pluckthun, A., A novel strategy to design binding molecules harnessing the modular nature of repeat proteins. *FEBS Lett* **2003,** *539* (1–3), 2–6.
18. Stumpp, M. T.; Forrer, P.; Binz, H. K.; Pluckthun, A., Designing repeat proteins: modular leucine-rich repeat protein libraries based on the mammalian ribonuclease inhibitor family. *J Mol Biol* **2003,** *332* (2), 471–87.

19. Kramer, L.; Renko, M.; Završnik, J.; Turk, D.; Seeger, M. A.; Vasiljeva, O.; Grütter, M. G.; Turk, V.; Turk, B., Non-invasive in vivo imaging of tumour-associated cathepsin B by a highly selective inhibitory DARPin. *Theranostics* **2017,** *7* (11), 2806-2821.
20. Goldstein, R.; Sosabowski, J.; Livanos, M.; Leyton, J.; Vigor, K.; Bhavsar, G.; Nagy-Davidescu, G.; Rashid, M.; Miranda, E.; Yeung, J.; Tolner, B.; Pluckthun, A.; Mather, S.; Meyer, T.; Chester, K., Development of the designed ankyrin repeat protein (DARPin) G3 for HER2 molecular imaging. *Eur J Nucl Med Mol Imaging* **2015,** *42* (2), 288–301.
21. Souied, E. H.; Devin, F.; Mauget-Faÿsse, M.; Kolář, P.; Wolf-Schnurrbusch, U.; Framme, C.; Gaucher, D.; Querques, G.; Stumpp, M. T.; Wolf, S., Treatment of Exudative Age-Related Macular Degeneration with a Designed Ankyrin Repeat Protein that Binds Vascular Endothelial Growth Factor: a Phase I/II Study. *Am J Ophthalmol* **2014,** *158* (4), 724–732.e2.
22. Fiedler, U.; Ekawardhani, S.; Cornelius, A.; Gilboy, P.; Bakker, T. R.; Dolado, I.; Stumpp, M. T.; Dawson, K. M., MP0250, a VEGF and HGF neutralizing DARPin(®) molecule shows high anti-tumor efficacy in mouse xenograft and patient-derived tumor models. *Oncotarget* **2017,** *8* (58), 98371-98383.
23. Pluckthun, A., Designed ankyrin repeat proteins (DARPins): binding proteins for research, diagnostics, and therapy. *Annu Rev Pharmacol Toxicol* **2015,** *55*, 489–511.
24. Zhang, Q.; Zmasek, C. M.; Godzik, A., Domain architecture evolution of pattern-recognition receptors. *Immunogenetics* **2010,** *62* (5), 263–72.
25. Han, B. W.; Herrin, B. R.; Cooper, M. D.; Wilson, I. A., Antigen Recognition by Variable Lymphocyte Receptors. *Science* **2008,** *321* (5897), 1834.
26. Bryant, C. E.; Monie, T. P., Mice, men and the relatives: cross-species studies underpin innate immunity. *Open Biol* **2012,** *2* (4), 120015.
27. Guo, P.; Hirano, M.; Herrin, B. R.; Li, J.; Yu, C.; Sadlonova, A.; Cooper, M. D., Dual nature of the adaptive immune system in lampreys. *Nature* **2009,** *459* (7248), 796-801.
28. Boehm, T.; McCurley, N.; Sutoh, Y.; Schorpp, M.; Kasahara, M.; Cooper, M. D., VLR-Based Adaptive Immunity. *Annu Rev Immunol* **2012,** *30* (1), 203-220.
29. Collins, B. C.; Gunn, R. J.; McKitrick, T. R.; Cummings, R. D.; Cooper, M. D.; Herrin, B. R.; Wilson, I. A., Structural Insights into VLR Fine Specificity for Blood Group Carbohydrates. *Structure (London, England : 1993)* **2017,** *25* (11), 1667-1678.e4.
30. Hong, X.; Ma, M. Z.; Gildersleeve, J. C.; Chowdhury, S.; Barchi, J. J.; Mariuzza, R. A.; Murphy, M. B.; Mao, L.; Pancer, Z., Sugar-Binding Proteins from Fish: Selection of High Affnity "Lambodies" That Recognize Biomedically Relevant Glycans. *ACS Chem. Biol.* **2013,** *8* (1), 152-160.
31. Luo, M.; Velikovsky, C. A.; Yang, X.; Siddiqui, M. A.; Hong, X.; Barchi, J. J.; Gildersleeve, J. C.; Pancer, Z.; Mariuzza, R. A., Recognition of the Thomsen-Friedenreich Pancarcinoma Carbohydrate Antigen by a Lamprey Variable Lymphocyte Receptor. *Journal of Biological Chemistry* **2013,** *288* (32), 23597-23606.
32. Gunn, R. J.; Herrin, B. R.; Acharya, S.; Cooper, M. D.; Wilson, I. A., VLR Recognition of TLR5 Expands the Molecular Characterization of Protein Antigen Binding by Non-Ig-based Antibodies. *J Mol Biol* **2018,** *430* (9), 1350-1367.
33. Velásquez, A. C.; Nomura, K.; Cooper, M. D.; Herrin, B. R.; He, S. Y., Leucine-rich-repeat-containing variable lymphocyte receptors as modules to target plant-expressed proteins. *Plant Methods* **2017,** *13* (1), 29.
34. Yu, C.; Liu, Y.; Chan, J. T. H.; Tong, J.; Li, Z.; Shi, M.; Davani, D.; Parsons, M.; Khan, S.; Zhan, W.; Kyu, S.; Grunebaum, E.; Campisi, P.; Propst, E. J.; Jaye, D. L.; Trudel, S.; Moran, M.

F.; Ostrowski, M.; Herrin, B. R.; Lee, F. E.-H.; Sanz, I.; Cooper, M. D.; Ehrhardt, G. R. A., Identification of human plasma cells with a lamprey monoclonal antibody. *JCI insight* **2016,** *1* (3), e84738.
35. Hwang, D.-E.; Ryou, J.-H.; Oh, J. R.; Han, J. W.; Park, T. K.; Kim, H.-S., Anti-Human VEGF Repebody Effectively Suppresses Choroidal Neovascularization and Vascular Leakage. *PloS One* **2016,** *11* (3), e0152522.
36. Zahnd, C.; Pecorari, F.; Straumann, N.; Wyler, E.; Plückthun, A., Selection and Characterization of Her2 Binding-designed Ankyrin Repeat Proteins. *Journal of Biological Chemistry* **2006,** *281* (46), 35167-35175.
37. Parker, R. N.; Grove, T. Z., Designing repeat proteins for biosensors and medical imaging. *Biochem Soc Trans* **2015,** *43* (5), 856-60.
38. Kummer, L.; Hsu, C.-W.; Dagliyan, O.; MacNevin, C.; Kaufholz, M.; Zimmermann, B.; Dokholyan, Nikolay V.; Hahn, Klaus M.; Plückthun, A., Knowledge-Based Design of a Biosensor to Quantify Localized ERK Activation in Living Cells. *Chem Biol* **2013,** *20* (6), 847–856.
39. Brient-Litzler, E.; Pluckthun, A.; Bedouelle, H., Knowledge-based design of reagentless fluorescent biosensors from a designed ankyrin repeat protein. *Protein Eng Des Sel* **2010,** *23* (4), 229–41.
40. Wang, Y.; Ballou, B.; Schmidt, B. F.; Andreko, S.; St. Croix, C. M.; Watkins, S. C.; Bruchez, M. P., Affibody-targeted fluorogen activating protein for in vivo tumor imaging. *Chemical Communications* **2017,** *53* (12), 2001-2004.
41. Chapman, A. M.; McNaughton, B. R., Resurfaced shape complementary proteins that selectively bind the oncoprotein gankyrin. *ACS Chem Biol* **2014,** *9* (10), 2223-8.
42. Gebauer, M.; Skerra, A., Engineered protein scaffolds as next-generation antibody therapeutics. *Curr Opin Chem Biol* **2009,** *13* (3), 245-55.
43. Geddes, K.; Magalhaes, J. G.; Girardin, S. E., Unleashing the therapeutic potential of NOD-like receptors. *Nat Rev Drug Discov* **2009,** *8* (6), 465-479.
44. Stumpp, M. T.; Binz, H. K.; Amstutz, P., DARPins: a new generation of protein therapeutics. *Drug Discov Today* **2008,** *13* (15–16), 695–701.
45. Cortajarena, A. L.; Yi, F.; Regan, L., Designed TPR modules as novel anticancer agents. *ACS Chem Biol* **2008,** *3* (3), 161–6.
46. Fiedler, M.; Skerra, A., *Non-Antibody Scaffolds as Alternative Therapeutic Agents*. 2014.
47. Main, E. R.; Jackson, S. E.; Regan, L., The folding and design of repeat proteins: reaching a consensus. *Curr Opin Struct Biol* **2003,** *13* (4), 482-9.
48. Andrade, M. A.; Perez-Iratxeta, C.; Ponting, C. P., Protein repeats: structures, functions, and evolution. *J Struct Biol* **2001,** *134* (2-3), 117-31.
49. Marcotte, E. M.; Pellegrini, M.; Yeates, T. O.; Eisenberg, D., A census of protein repeats. *J Mol Biol* **1999,** *293* (1), 151–60.
50. Lee, S.-C.; Park, K.; Han, J.; Lee, J.-j.; Kim, H. J.; Hong, S.; Heu, W.; Kim, Y. J.; Ha, J.-S.; Lee, S.-G.; Cheong, H.-K.; Jeon, Y. H.; Kim, D.; Kim, H.-S., Design of a binding scaffold based on variable lymphocyte receptors of jawless vertebrates by module engineering. *Proceedings of the National Academy of Sciences* **2012,** *109* (9), 3299.
51. Porebski, B. T.; Buckle, A. M., Consensus protein design. *Protein Engineering, Design and Selection* **2016,** *29* (7), 245-251.
52. Uehling, J.; Deveau, A.; Paoletti, M., Do fungi have an innate immune response? An NLR-based comparison to plant and animal immune systems. *PLOS Pathogens* **2017,** *13* (10), e1006578.

53. Ausubel, F. M., Are innate immune signaling pathways in plants and animals conserved? *Nat Immunol* **2005,** *6* (10), 973–9.
54. Meunier, E.; Broz, P., Evolutionary Convergence and Divergence in NLR Function and Structure. *Trends in Immunology* **2017,** *38* (10), 744-757.
55. Urbach, J. M.; Ausubel, F. M., The NBS-LRR architectures of plant R-proteins and metazoan NLRs evolved in independent events. *Proceedings of the National Academy of Sciences* **2017,** *114* (5), 1063.
56. Proell, M.; Riedl, S. J.; Fritz, J. H.; Rojas, A. M.; Schwarzenbacher, R., The Nod-like receptor (NLR) family: a tale of similarities and differences. *PloS One* **2008,** *3* (4), e2119.
57. Akira, S., Toll-like receptors: lessons from knockout mice. *Biochem Soc Trans* **2000,** *28* (5), 551–556.
58. Hayashi, F.; Smith, K. D.; Ozinsky, A.; Hawn, T. R.; Yi, E. C.; Goodlett, D. R.; Eng, J. K.; Akira, S.; Underhill, D. M.; Aderem, A., The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. *Nature* **2001,** *410* (6832), 1099–1103.
59. Janeway, C. A., Jr.; Medzhitov, R., Innate immune recognition. *Annu Rev Immunol* **2002,** *20*, 197–216.
60. Akira, S.; Takeda, K., Functions of Toll-like receptors: lessons from KO mice. *C R Biol* **2004,** *327* (6), 581–589.
61. Xu, D.; Liu, H.; Komai-Koma, M., Direct and indirect role of Toll-like receptors in T cell mediated immunity. *Cell Mol Immunol* **2004,** *1* (4), 239–46.
62. Alder, M. N.; Rogozin, I. B.; Iyer, L. M.; Glazko, G. V.; Cooper, M. D.; Pancer, Z., Diversity and Function of Adaptive Immune Receptors in a Jawless Vertebrate. *Science* **2005,** *310* (5756), 1970.
63. Kim, H. M.; Park, B. S.; Kim, J.-I.; Kim, S. E.; Lee, J.; Oh, S. C.; Enkhbayar, P.; Matsushima, N.; Lee, H.; Yoo, O. J.; Lee, J.-O., Crystal Structure of the TLR4-MD-2 Complex with Bound Endotoxin Antagonist Eritoran. *Cell* **2007,** *130* (5), 906–917.
64. Oosting, M.; Cheng, S. C.; Bolscher, J. M.; Vestering-Stenger, R.; Plantinga, T. S.; Verschueren, I. C.; Arts, P.; Garritsen, A.; van Eenennaam, H.; Sturm, P.; Kullberg, B. J.; Hoischen, A.; Adema, G. J.; van der Meer, J. W.; Netea, M. G.; Joosten, L. A., Human TLR10 is an anti-inflammatory pattern-recognition receptor. *Proc Natl Acad Sci U S A* **2014,** *111* (42), E4478–84.
65. Rosenstiel, P.; Jacobs, G.; Till, A.; Schreiber, S., NOD-like receptors: ancient sentinels of the innate immune system. *Cell Mol Life Sci* **2008,** *65* (9), 1361–77.
66. Parker, R.; Mercedes-Camacho, A.; Grove, T. Z., Consensus design of a NOD receptor leucine rich repeat domain with binding affinity for a muramyl dipeptide, a bacterial cell wall fragment. *Protein Sci* **2014,** *23* (6), 790–800.
67. Kufer, T. A.; Banks, D. J.; Philpott, D. J., Innate immune sensing of microbes by Nod proteins. *Ann N Y Acad Sci* **2006,** *1072*, 19–27.
68. Kaparakis, M.; Philpott, D. J.; Ferrero, R. L., Mammalian NLR proteins; discriminating foe from friend. *Immunol Cell Biol* **2007,** *85* (6), 495–502.
69. Zhao, Y.; Yang, J.; Shi, J.; Gong, Y.-N.; Lu, Q.; Xu, H.; Liu, L.; Shao, F., The NLRC4 inflammasome receptors for bacterial flagellin and type III secretion apparatus. *Nature* **2011,** *477* (7366), 596–600.
70. Hamada, M.; Shoguchi, E.; Shinzato, C.; Kawashima, T.; Miller, D. J.; Satoh, N., The complex NOD-like receptor repertoire of the coral Acropora digitifera includes novel domain combinations. *Mol Biol Evol* **2013,** *30* (1), 167–76.

71. Lange, C.; Hemmrich, G.; Klostermeier, U. C.; Lopez-Quintero, J. A.; Miller, D. J.; Rahn, T.; Weiss, Y.; Bosch, T. C.; Rosenstiel, P., Defining the origins of the NOD-like receptor system at the base of animal evolution. *Mol Biol Evol* **2011,** *28* (5), 1687–702.
72. Yuen, B.; Bayes, J. M.; Degnan, S. M., The characterization of sponge NLRs provides insight into the origin and evolution of this innate immune gene family in animals. *Mol Biol Evol* **2014,** *31* (1), 106–20.
73. Waters, E. A.; Shusta, E. V., The variable lymphocyte receptor as an antibody alternative. *Curr Opin Biotechnol* **2018,** *52*, 74-79.
74. Herrin, B. R.; Alder, M. N.; Roux, K. H.; Sina, C.; Ehrhardt, G. R. A.; Boydston, J. A.; Turnbough, C. L.; Cooper, M. D., Structure and specificity of lamprey monoclonal antibodies. *Proceedings of the National Academy of Sciences* **2008,** *105* (6), 2040.
75. Marold, Jacob D.; Kavran, Jennifer M.; Bowman, Gregory D.; Barrick, D., A Naturally Occurring Repeat Protein with High Internal Sequence Identity Defines a New Class of TPR-like Proteins. *Structure* **2015,** *23* (11), 2055-2065.
76. Lee, J. S.; Kim, J.; Im, S. P.; Kim, S. W.; Lazarte, J. M. S.; Jung, J. W.; Gong, T. W.; Kim, Y. R.; Lee, J. H.; Kim, H. J.; Jung, T. S., Generation and characterization of hagfish variable lymphocyte receptor B against glycoprotein of viral hemorrhagic septicemia virus (VHSV). *Molecular Immunology* **2018,** *99*, 30-38.
77. Velikovsky, C. A.; Deng, L.; Tasumi, S.; Iyer, L. M.; Kerzic, M. C.; Aravind, L.; Pancer, Z.; Mariuzza, R. A., Structure of a lamprey variable lymphocyte receptor in complex with a protein antigen. *Nature Structural &Amp; Molecular Biology* **2009,** *16*, 725.
78. Kirchdoerfer, R. N.; Herrin, B. R.; Han, B. W.; Turnbough, C. L., Jr.; Cooper, M. D.; Wilson, I. A., Variable lymphocyte receptor recognition of the immunodominant glycoprotein of Bacillus anthracis spores. *Structure (London, England : 1993)* **2012,** *20* (3), 479-486.
79. Wezner-Ptasińska, M.; Krowarsch, D.; Otlewski, J., Design and characteristics of a stable protein scaffold for specific binding based on variable lymphocyte receptor sequences. *BBA-Proteins Proteom* **2011,** *1814* (9), 1140-1145.
80. Lee, S.-C.; Park, K.; Han, J.; Lee, J.-j.; Kim, H. J.; Hong, S.; Heu, W.; Kim, Y. J.; Ha, J.-S.; Lee, S.-G.; Cheong, H.-K.; Jeon, Y. H.; Kim, D.; Kim, H.-S., Design of a binding scaffold based on variable lymphocyte receptors of jawless vertebrates by module engineering. *Proc Natl Acad Sci U S A* **2012,** *109* (9), 3299-3304.
81. Yun, M.; Kim, D. Y.; Lee, J. J.; Kim, H. S.; Kim, H. S.; Pyo, A.; Ryu, Y.; Kim, T. Y.; Zheng, J. H.; Yoo, S. W.; Hyun, H.; Oh, G.; Jeong, J.; Moon, M.; Min, J. H.; Kwon, S. Y.; Kim, J. Y.; Chung, E.; Hong, Y.; Lee, W.; Kim, H. S.; Min, J. J., A High-Affinity Repebody for Molecular Imaging of EGFR-Expressing Malignant Tumors. *Theranostics* **2017,** *7* (10), 2620-2633.
82. Lee, J.-j.; Kim, H. J.; Yang, C.-S.; Kyeong, H.-H.; Choi, J.-M.; Hwang, D.-E.; Yuk, J.-M.; Park, K.; Kim, Y. J.; Lee, S.-G.; Kim, D.; Jo, E.-K.; Cheong, H.-K.; Kim, H.-S., A High-Affinity Protein Binder that Blocks the IL-6/STAT3 Signaling Pathway Effectively Suppresses Non–Small Cell Lung Cancer. *Molecular Therapy* **2014,** *22* (7), 1254-1265.
83. Wezner-Ptasinska, M.; Otlewski, J., Selection of specific interactors from phage display library based on sea lamprey variable lymphocyte receptor sequences. *BBA-Proteins Proteom* **2015,** *1854* (12), 1833-1841.
84. Lee, J. S.; Kim, J.; Im, S. P.; Kim, S. W.; Jung, J. W.; Lazarte, J. M. S.; Lee, J.-H.; Thompson, K. D.; Jung, T. S., Expression and characterization of monomeric variable lymphocyte receptor B specific to the glycoprotein of viral hemorrhagic septicemia virus (VHSV). *J Immunol Methods* **2018,** *462*, 48-53.

85. Kim, H.-Y.; Lee, J.-j.; Kim, N.; Heo, W. D.; Kim, H.-S., Tracking protein–protein interaction and localization in living cells using a high-affinity molecular binder. *Biochemical and Biophysical Research Communications* **2016,** *470* (4), 857-863.

Chapter 4. *Hydra Magnipapillata* Innate Immune 42PRs

Jennifer P. McCord, V. Grey Fritz, Tijana Z. Grove

Department of Chemistry, Virginia Tech, Blacksburg, VA 24061

Keywords: innate immune receptors, pattern recognition receptors, tetratricopeptide repeats

4.1 Abstract

Repeat proteins are instrumental in a wide variety of biological processes, including the recognition of pathogen-associated molecular patterns (PAMPs) by the innate immune system. The natural molecular diversity of PAMPs make innate immune proteins an interesting starting point for the design of alternative binding proteins, whose utility is often restricted to protein and peptide ligands. Recently, reports emerged of early diverging organisms with innate immune systems that differ from the well-studied mammalian innate immune proteins in that they use different repeat domains for recognition. The *Hydra magnipapillata* innate immune system has a particularly large repertoire of genes containing tetratricopeptide repeat (TPR) domains. We found that these genes contained particularly long contiguous repeat domains that had 42 amino acids per repeat with high sequence similarity, which is characteristic of repeat proteins that bind nucleic acid ligands. We undertook statistical design to create a binding protein based on the *H. magnipapillata* innate immune TPR proteins in an effort to make a sensor to bind nucleic acid, glycan and glycoprotein ligands.

4.2 Introduction

Since the 1960s when the first polyclonal antibodies were produced by injecting rabbits with antigens of interest, immunoglobulin scaffolds have been used as a workhorse for specific

molecular recognition of targets of interest in biomedicine, research, and industry. However, while there are nearly half a million antibodies on the market, many for research applications are not specific binders for their targets.[1] Additionally, their large size (~150 kDa for IgG antibodies) and reliance on disulfide bonds for stability makes them unsuitable for applications such as the reducing environment of the cytoplasm.[2] Due to evolutionary pressures, mammalian immunoglobulins are also poor binders for carbohydrates of biomedical interest, such as those responsible for blood typing.[3]

As a result, researchers in the past several decades have proposed a wide array of alternative binding scaffolds with the goal to improve the selectivity, specificity, and cost required for binding proteins, with much success.[4-9] Repeat proteins are particularly promising, as their modularity makes them straightforward to design and adaptable for different targets.[10-11] One drawback of alternative scaffolds is that they often are restricted to protein and peptide targets,[10, 12-13] leaving out many classes of biologically relevant molecules. However, repeat proteins from the immune system often retain their ability to bind non-protein targets.[3, 14] Our lab previously engineered a leucine-rich repeat (LRR) protein based on the recognition domain of mammalian nod-like receptor proteins stereoselectively bound muramyl dipeptide with micromolar affinity.[15]

Researchers looking into the evolution of the innate immune system found genes throughout the biosphere that look much like the well-characterized nod-like receptors in mammals, except that they contain repeat domains other than LRR domains.[16-18] A freshwater polyp, *Hydra magnipapillata*, contains a particularly large repertoire of nod-like receptor-like genes that have tetratricopeptide repeats (TPRs) instead of LRRs.

This was quite interesting as TPR domains typically have protein ligands,[19-23] while the innate immune system is responsible for the recognition of a number of molecularly diverse ligands.[24-34] Since these *H. magnipapillata* genes appeared to belong to the innate immune system, we hypothesized that they could potentially bind non-protein targets. As we began aligning the TPRs in these genes, we observed conserved 8-residue "gaps" between predicted TPR sequences. Therefore, we included these residues in our definition of a repeat, making a TPR that was 42 amino acids in length, termed 42PRs by Marold et al.[19] TPRs are typically 34 amino acid residues in length, and TPRs of this length have been studied extensively.[20-22, 35-41] Recently, there have been efforts to better characterize TPR-like repeat domains of different length, such as the pentatricopeptide repeat (PPR) common in the plant kingdom,[42-44] and the 42PRs from *P. anserina* (*Pa*).[19] An interesting characteristic of PPRs, hypothesized to be shared by 42PRs based on high sequence similarity and number of contiguous repeats, is their ability to binding nucleic acid ligands in addition to protein ligands. The 42PRs from *H. magnipapillata* could show similarly diverse ligand binding capabilities, and set out to produce a protein based on these 42PRs that would bind to biologically relevant targets, particularly lipopolysaccharide, as that ligand is a demonstrated antigen to *H. magnipapillata*.[17]

4.3 Materials and Methods

4.3.1 Consensus Design and Multiple Sequence Alignment

The SMART database in genomic mode was searched for genes in *Hydra vulgaris* containing both NB-ARC and TPR domains. These TPR domains were searched using the BLAST tool in the NCBI *Hydra vulgaris* genome library, and those genes identified by Lange, et al.[16] as containing both NB-ARC and TPR domains were translated using the Bioinformatics Sequence Manipulation Suite.[45] TPR domains were manually extracted to Microsoft Excel. We

noticed a gap of 8 amino acids in between each "TPR" repeat, and therefore decided to redefine our repeats to include the 42-amino acid motif, termed 42PRs by the Barrick lab.[19] For our first generation of TPRs, we designed a strict consensus sequence, modifying the conserved cysteine at position 15 to an alanine for a cysteine-free scaffold as CTPR.[41] Characterization of these proteins revealed that they were soluble but unfolded, so we designed a second generation. We identified hydrophobic residues (methionine at position 19 and isoleucine at position 42) that were predicted to be solvent-exposed. We mutated them to the next most common hydrophilic residues for those positions (threonine in both cases). These proteins were soluble, but unfolded. We next identified a conserved salt bridge (between the A and B helices, the lysine at position 9 and the aspartate at position 33) which in the CTPR family of proteins was a hydrophobic interaction between tryptophan and leucine, and modified our original consensus sequence to include this interaction.[41]

4.3.2 Gene Synthesis and Cloning

Enzymes were purchased from New England Biolabs (Ipswich, MA). For cloning of the 42PR protein, genes were designed consisting of a two-helix repeat and a capping helix. Capping helices have been shown to increase the stability of TPR proteins.[41] The capping helix was separated from the two-helix repeat by a BglII restriction site. First-generation genes were synthesized by Klenow extension of four overlapping synthetic oligonucleotides coding for the gene essentially as published previously.[46] Briefly, 5 µL 100 mM overlapping primer pairs were diluted to 45 µL in annealing buffer consisting of 10 mM Tris HCl, 50 mM NaCl, and 6 mM $MgCl_2$. The oligonucleotides were then annealed by boiling for 30 minutes in a 1 L water bath, that was allowed to cool to room temperature. Klenow DNA polymerase and 1 µL each dNTP were added, and allowed to react for 30 mins at 37 °C. Synthesis of the full gene was completed

by a two-step PCR process using Taq DNA polymerase; the first step was 5 rounds of PCR without amplification primers to encourage annealing between overlapping Klenow products; the second was 35 rounds of PCR with amplification primers encoding for restriction sites. Annealing temperatures were determined by New England Biolabs T_m calculator. PCR products were worked up before ligation into plasmid *pProExHtam* as below.

Second-generation synthetic genes were synthesized by GENEWIZ (South Plainfield, NJ) and cloned into plasmid *pProExHtam* by ligation of restrictions sites BamHI and HindIII. Recursive ligation to achieve repeat proteins of the desired length was achieved by double digestion of the gene with BamHI and HindIII, which was cloned into a linearized dephosphorylated plasmid. Gene identity was confirmed by sequencing at the Bioinformatics Institute of Virginia Tech.

4.3.3 Protein Expression and Purification

Overnight cultures of BL21 (DE3) cells were diluted 1:100 in 1 L of Luria broth media at 37°C, with shaking at 180 rpm, and were grown to an OD_{600} of 0.5–0.8. Expression was induced with 0.5 mM IPTG, followed by 16 h of expression at 17 °C. The cells were harvested by centrifugation at 5,000 rpm for 15 min and the pellets were resuspended in lysis buffer consisting of 50 mM Tris, 300 mM sodium chloride, and 0.1% Tween 20, pH 8 and frozen at −80 °C until purification. To purify proteins, the cell pellet was thawed and sonified for 30 s at 30% power using a microtip and Mison sonifier. Lysed cells were then centrifuged at 16,000 rpm for 30 min; the supernatant was discarded. Lysis buffer with 8 M urea was added to the cell pellet and the suspension was centrifuged again at 16,000 RPM for 30 min; the protein supernatant was collected. Proteins were purified using Ni-NTA affinity purification protocol under denaturing conditions and eluted with 300 mM imidazole in lysis buffer. Proteins were then further purified

on the size-exclusion column in 50 mM sodium phosphate buffer pH 8 with 150 mM NaCl. Proteins were quantified by absorption at 280 nm using an extinction coefficient of 27, 960 M^{-1} cm^{-1}, calculated from the amino acid sequence using the Expasy Protparam tool.[47]

4.3.4 Size Exclusion Chromatography

Akta Prime Plus FPLC was used for size exclusion chromatography. Further purification was completed on the Superdex 75 16/600 Prep Grade column in 150 m*M* sodium chloride and 50 m*M* sodium phosphate buffer pH 8 at a flow rate of 1 ml/min. The Superdex 75 10/300 analytical column was used for analysis of molecular weights under the same conditions. A comparison to known standards (Bio-Rad) allowed for determination of the molecular weights and oligomeric states of each 42PR protein.

4.3.5 Circular Dichroism

CD spectra were acquired using 5–10 µ*M* protein samples in 10 m*M* phosphate buffer pH 7.4 with 10 mM NaCl using a Jasco J-815 CD spectrometer. Far-UV CD (190–260 nm) spectra were recorded at 25 °C to assess the secondary structure of CLRR2. Each sample was recorded three times, from 190 to 260 nm in a 2 mm pathlength cuvette, and averaged. Data collected using a 1 nm bandwidth, 2 nm data pitch, and a data integration time of 1 s, was normalized to units of mean residue ellipticity for all samples.

4.4 Results and Discussion

4.4.1 *Hydra magnipapillata* NLR 42PRs

The *H. magnipapillata* NLR 42PRs share structural features both with TPRs[22, 37, 41] and 42PRs.[19] The TPR repeat is characterized by a helix-turn-helix motif, repeated typically three

times in sequence.[22] 42PRs differ in helical length; they have one additional helical turn in each helix, so generally each position in a 42PR matches the same position in TPRs, plus 5. We will use the number determined for TPRs here, to match with other literature. Characteristic structure-determining residues shared among TPRs and 42PRs are the alanine or glycine at position 8; a large aliphatic or aromatic amino acid at position 11; glycine at position 15; alanine at position 20; leucine, phenylalanine or tyrosine at position 24; alanine at position 27; leucine at position 28; and proline at position 32 (which becomes position 3 in 42PRs). The multiple sequence alignments for each type of repeat are visualized by weblogo[48] in **Figure 4.1a.**

Like *P. Anisera*,[19] whose 42PRs are found in repeat domains consisting of 16 repeats, the 42PRs of *H. magnipapillata* are found in much longer sequences of repeats than most TPRs, which tend to be found in 3-repeat clusters.[22] The 42PRs of *H. magnipapillata* are found in longer segments, with the most common number of repeats per gene being 9 (**Figure 4.1b**), and a significant number having 24 or more contiguous repeats.

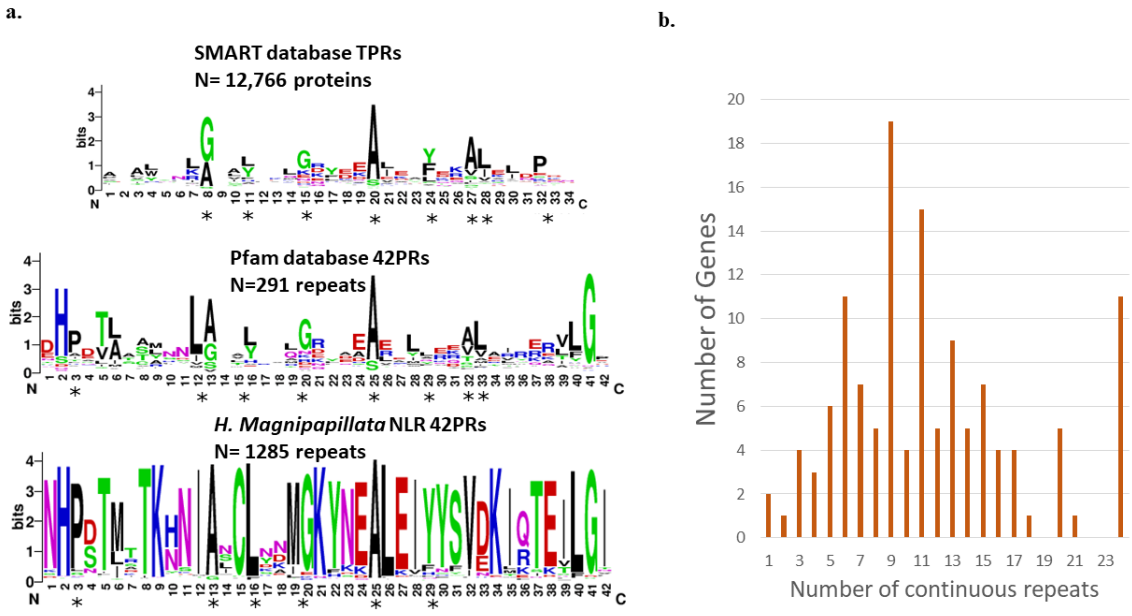


Figure 4.1. a. Mutual sequence alignment of all TPRs (top), 42PRs from the Pfam database (middle), and *Hydra* NLR 42PRs (bottom). b. Number of continuous repeats in *Hydra* NLR 42PR genes.

4.4.2 Consensus Design

Consensus design is a technique that takes the most common amino acid at each position from a library of homologous proteins. The general idea is that among homologous proteins, the most common amino acid will be the most stabilizing while maintaining function. Studies of several enzymes and antibodies have demonstrated that mutating a random residue in a protein with the consensus residue at that position is stabilizing about half the time.[49] This is likely because not all residues contribute equally to the protein structure. Repeat proteins have particularly pronounced differences between important, structure-determining residues and hypervariable residues, which are not conserved in multiple sequence alignments and are responsible for the specific interactions between protein and ligand.[50-51]

4.4.3 Protein Expression and Characterization

We expressed our first generation protein as the consensus sequence, mutating the cysteine at position at position 15 to alanine to avoid unfavorable disulfide bond formation, with an added extra helix at the end as this is known to be stabilizing for TPR proteins.[22] We chose to look at the 3-repeat protein, as TPRs of fewer repeats are often unstable, but smaller scaffolds are more favorable for binding applications.[6, 9, 52] This protein was insoluble, and did not refold (**Figure 4.2a)**.

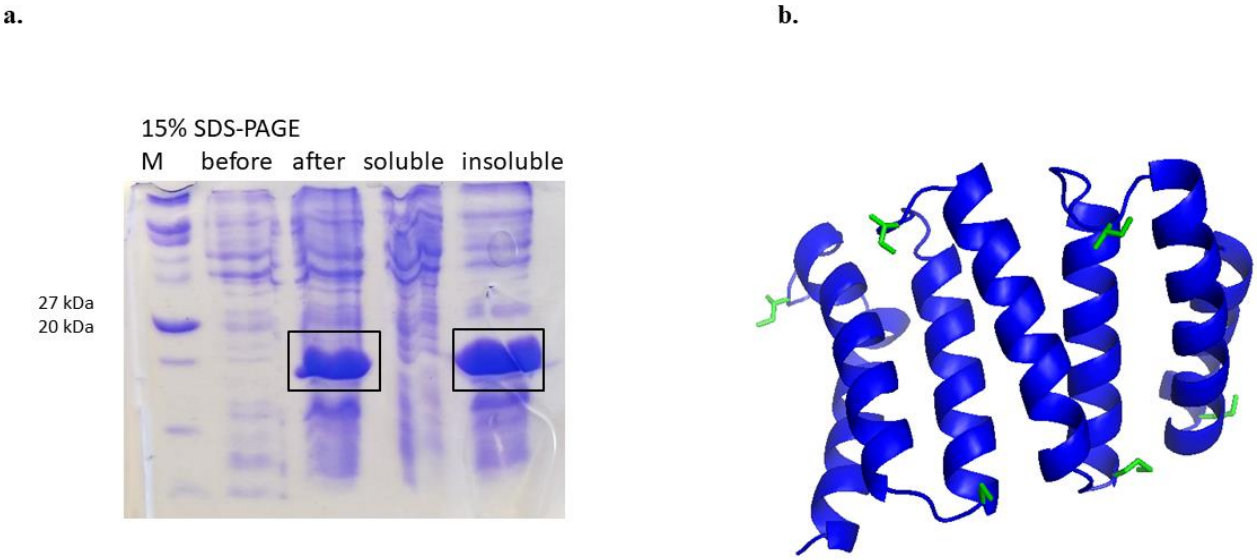


Figure 4.2 a. SDS-PAGE gel expression of the 3-repeat consensus sequence protein. All overexpressed protein is in the insoluble fraction. b. MUSTER model of the 3-repeat consensus sequence protein. Solvent-exposed hydrophobic residues are highlighted in green.

4.4.4 Protein Redesign

Examination of the MUSTER model of our 3-repeat consensus protein (**Figure 4.2b**) revealed there were two predicted solvent-exposed hydrophobic residues per repeat (methionine at position 19, and isoleucine at position 42). We therefore mutated these residues to the most common hydrophilic residue in these positions (threonine in both cases).

We found that this protein expressed well, at 10 mg/L, and eluted in a single symmetrical peak on the size-exclusion column (**Figure 4.3a**). From a calibration curve of known protein standards (**SI Figure 4.2**), this protein appeared to be 40 kDa, as opposed to its expected size of 20 kDa. Circular dichroism spectra (**Figure 4.3b**) revealed a primarily random coil protein structure, which would make the protein appear larger on the size-exclusion column. To look at the propensity of the protein to form helical structures, we looked at the CD spectra in 40% trifluoroethanol, which is known to increase the secondary structure of proteins. Upon this addition, we saw a classic α-helical spectra, with characteristic bands at 208 and 220 nm. This

suggested to us that the protein was unfolded, but had the propensity to form alpha helices. We therefore decided to redesign the protein to increase its folding.

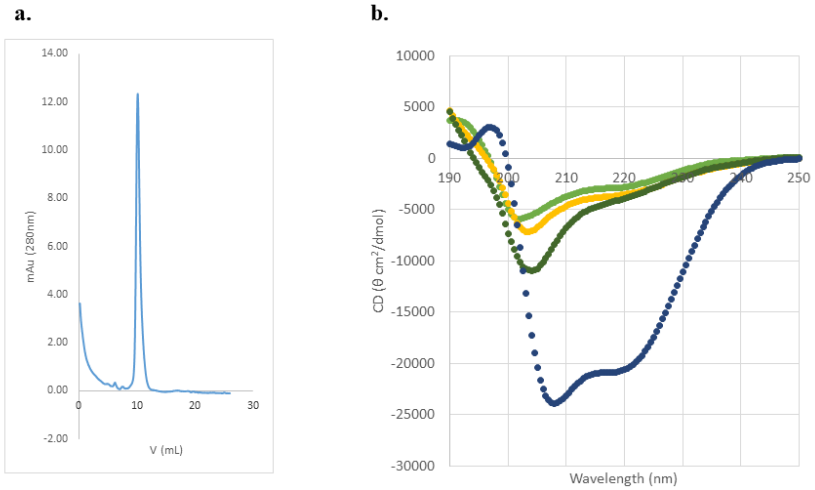


Figure 4.3. a. Size-exclusion chromatography trace. Expected protein size is 20 kDa; size as calculated by a calibration curve of known proteins is 40 kDa. b. Circular dichroism spectra of first-generation Hy42PR. Light green trace is immediately following dialysis to remove urea; yellow trace is following refolding on the size-exclusion column; dark green trace is following refolding on the Ni-NTA column; dark blue trace is in 40% trifluoroethanol.

4.4.5 Second Protein Redesign

To address the lack of helical character, we compared the sequences and predicted crystal structure of *H. magnipapillata* NLR 42PRs with that of highly stable scaffold CTPR.[41] A major difference we found was an apparent salt bridge between the lysine at position 9 and the aspartate at position 33 in Hy42PR (**Figure 4.4**) which was a hydrophobic packing interaction between tryptophan and leucine in the equivalent positions in CTPR. We hypothesize this difference could be responsible for the lack of helical secondary structure in Hy42PR3, as an excess of hydrophilic residues could minimize hydrophobic exclusion that stabilizes the folded state of the protein. We therefore modified these residues to the tryptophan and leucine that are found in CTPR.

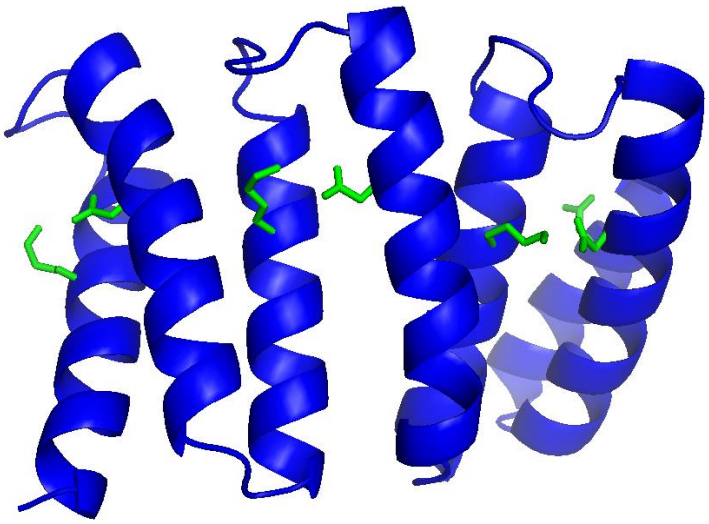


Figure 4.4. MUSTER model of first-generation 3-repeat Hy42PR3. Lysine at position 9 and aspartate at position 33 are highlighted in green.

4.4.5 Second Generation Protein Expression and Characterization

For this generation of proteins, we found that the 3-repeat protein expressed in the inclusion bodies and was resistant to refolding by dialysis, on the Ni-NTA column, and on the size-exclusion column. We decided to extend this protein to 5 repeats, given its similarity to *Pa* 42PRs. Unfortunately, we found that the 5-repeat protein was likewise insoluble (**Figure 4.5**) and resistant to folding by the above methods.

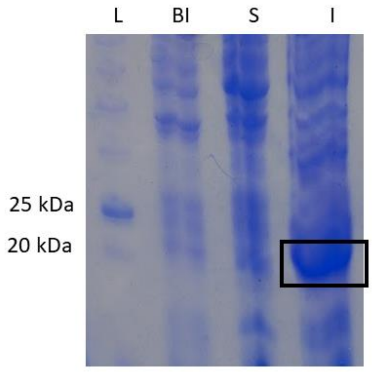


Figure 4.5 SDS-PAGE of second-generation Hy42PR5.

The high sequence similarity of *H. magnipapillata* 42PRs, while very interesting from an evolutionary perspective, makes statistical design challenging when there is a lack of stability

with *E. coli* expression. Since each repeat is so similar, it is difficult to determine which combination of residues is responsible for the lack of solubility and folding. These problems could likely be fixed by increasing the diversity of the multiple sequence alignment by a BLAST search[53] that did not take organism or function of the repeat into account. The binding site of the *H. magnipapillata* NLRs could be grafted onto the resulting scaffold for a sensor that bound physiologically relevant ligands.

4.5 Conclusions

Hydra magnipapillata has a particularly large repertoire of genes that contain domains characteristic of classic innate immunity, and contain TPR repeats rather than LRRs. We set out to investigate the structural properties of these TPRs, and found that they contain an unusual 42 amino acids, along with high sequence similarity and tend to be found in relatively very large domains. This high sequence conservation and length indicates the possibility of binding polyvalent targets such as that found in pathogenic DNA.[54] While the proteins we have designed based on these unusual 42PRs have had issues with solubility, this study has given us a better understanding of the features of the *H. magnipapillata* innate immune protein sensors. Better knowledge of these features can lead to better rational design of pathogenic sensors.

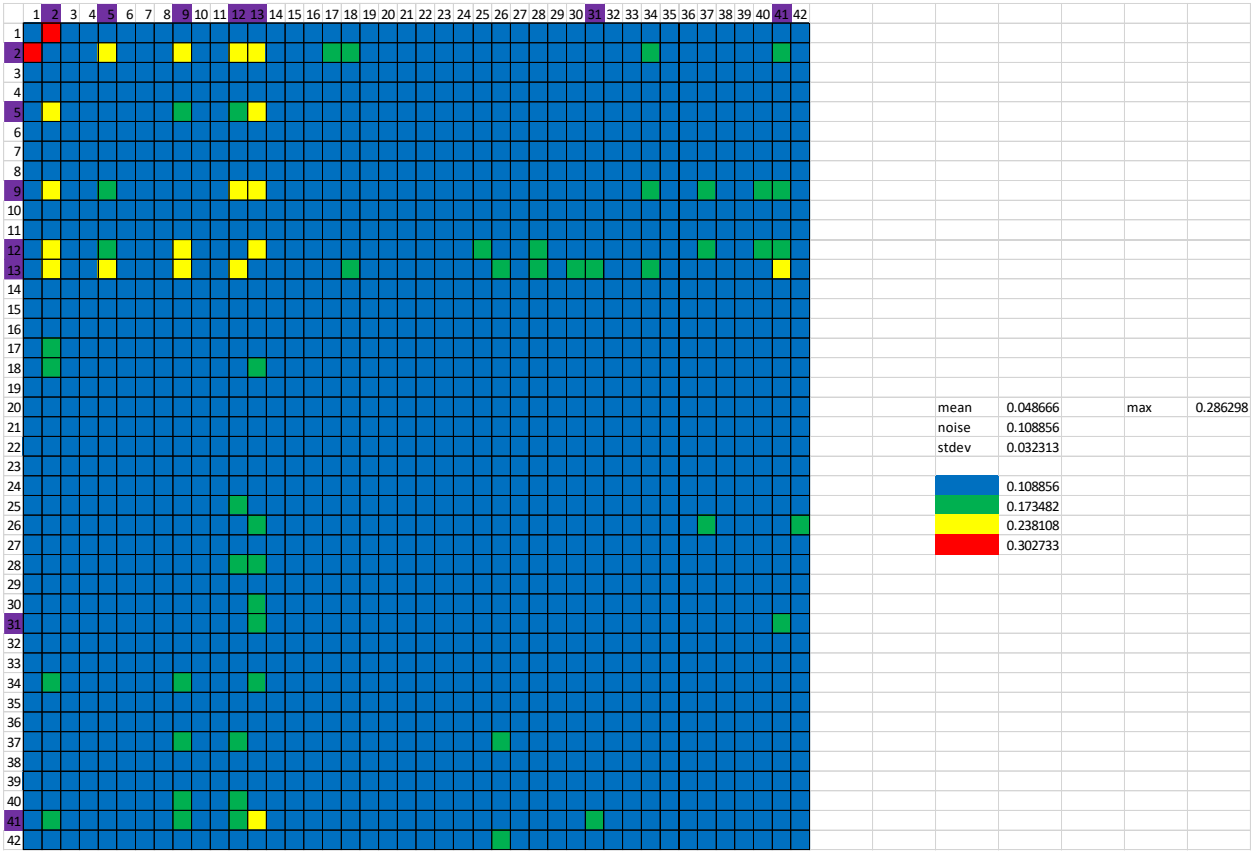
4.6 References

1. Berglund, L.; Björling, E.; Oksvold, P.; Fagerberg, L.; Asplund, A.; Al-Khalili Szigyarto, C.; Persson, A.; Ottosson, J.; Wernérus, H.; Nilsson, P.; Lundberg, E.; Sivertsson, Å.; Navani, S.; Wester, K.; Kampf, C.; Hober, S.; Pontén, F.; Uhlén, M., A Genecentric Human Protein Atlas for Expression Profiles Based on Antibodies. *Mol Cell Proteomics* **2008,** *7* (10), 2019–2027.
2. Bradbury, A.; Pluckthun, A., Reproducibility: Standardize antibodies used in research. *Nature* **2015,** *518* (7537), 27-29.
3. Collins, B. C.; Gunn, R. J.; McKitrick, T. R.; Cummings, R. D.; Cooper, M. D.; Herrin, B. R.; Wilson, I. A., Structural Insights into VLR Fine Specificity for Blood Group Carbohydrates. *Structure (London, England : 1993)* **2017,** *25* (11), 1667-1678.e4.
4. Könning, D.; Kolmar, H., Beyond antibody engineering: directed evolution of alternative binding scaffolds and enzymes using yeast surface display. *Microbial Cell Factories* **2018,** *17* (1), 32.
5. Grönwall, C.; Ståhl, S., Engineered affinity proteins—Generation and applications. *J Biotechnol* **2009,** *140* (3–4), 254–269.
6. Skerra, A., Alternative non-antibody scaffolds for molecular recognition. *Curr Opin Biotechnol* **2007,** *18* (4), 295-304.
7. Hosse, R. J.; Rothe, A.; Power, B. E., A new generation of protein display scaffolds for molecular recognition. *Protein Sci* **2006,** *15* (1), 14–27.
8. Binz, H. K.; Pluckthun, A., Engineered proteins as specific binding reagents. *Curr Opin Biotechnol* **2005,** *16* (4), 459–469.
9. Nygren, P. A.; Skerra, A., Binding proteins from alternative scaffolds. *J Immunol Methods* **2004,** *290* (1–2), 3–28.
10. Parker, R. N.; Grove, T. Z., Designing repeat proteins for biosensors and medical imaging. *Biochem Soc Trans* **2015,** *43* (5), 856-860.
11. Main, E. R.; Lowe, A. R.; Mochrie, S. G.; Jackson, S. E.; Regan, L., A recurring theme in protein engineering: the design, stability and folding of repeat proteins. *Curr Opin Struct Biol* **2005,** *15* (4), 464-471.
12. Fiedler, M.; Skerra, A., Non-Antibody Scaffolds as Alternative Therapeutic Agents. In *Handbook of Therapeutic Antibodies*, Wiley-VCH Verlag GmbH & Co. KGaA: 2014; pp 435-474.
13. Boersma, Y. L.; Pluckthun, A., DARPins and other repeat protein scaffolds: advances in engineering and applications. *Curr Opin Biotechnol* **2011,** *22* (6), 849-857.
14. Waters, E. A.; Shusta, E. V., The variable lymphocyte receptor as an antibody alternative. *Curr Opin Biotechnol* **2018,** *52*, 74-79.
15. Parker, R.; Mercedes-Camacho, A.; Grove, T. Z., Consensus design of a NOD receptor leucine rich repeat domain with binding affinity for a muramyl dipeptide, a bacterial cell wall fragment. *Protein Sci* **2014,** *23* (6), 790–800.
16. Lange, C.; Hemmrich, G.; Klostermeier, U. C.; Lopez-Quintero, J. A.; Miller, D. J.; Rahn, T.; Weiss, Y.; Bosch, T. C.; Rosenstiel, P., Defining the origins of the NOD-like receptor system at the base of animal evolution. *Mol Biol Evol* **2011,** *28* (5), 1687–1702.
17. Bosch, T. C. G.; Augustin, R.; Anton-Erxleben, F.; Fraune, S.; Hemmrich, G.; Zill, H.; Rosenstiel, P.; Jacobs, G.; Schreiber, S.; Leippe, M.; Stanisak, M.; Grötzinger, J.; Jung, S.; Podschun, R.; Bartels, J.; Harder, J.; Schröder, J.-M., Uncovering the evolutionary history of

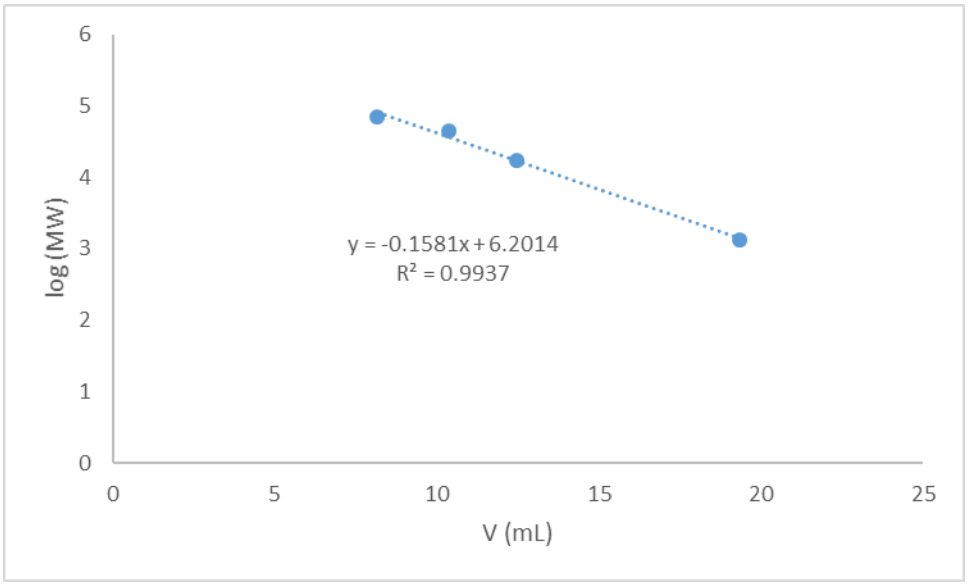
innate immunity: The simple metazoan Hydra uses epithelial cells for host defence. *Developmental & Comparative Immunology* **2009,** *33* (4), 559-569.
18. Hamada, M.; Shoguchi, E.; Shinzato, C.; Kawashima, T.; Miller, D. J.; Satoh, N., The complex NOD-like receptor repertoire of the coral Acropora digitifera includes novel domain combinations. *Mol Biol Evol* **2013,** *30* (1), 167–176.
19. Marold, Jacob D.; Kavran, Jennifer M.; Bowman, Gregory D.; Barrick, D., A Naturally Occurring Repeat Protein with High Internal Sequence Identity Defines a New Class of TPR-like Proteins. *Structure* **2015,** *23* (11), 2055-2065.
20. Cortajarena, A. L.; Regan, L., Ligand binding by TPR domains. *Protein Sci* **2006,** *15* (5), 1193-1198.
21. Cortajarena, A. L.; Kajander, T.; Pan, W.; Cocco, M. J.; Regan, L., Protein design to understand peptide ligand recognition by tetratricopeptide repeat proteins. *Protein Eng Des Sel* **2004,** *17* (4), 399-409.
22. D'Andrea, L. D.; Regan, L., TPR proteins: the versatile helix. *Trends Biochem Sci* **2003,** *28* (12), 655–662.
23. Blatch, G. L.; Lassle, M., The tetratricopeptide repeat: a structural motif mediating protein-protein interactions. *BioEssays* **1999,** *21* (11), 932–939.
24. Gunn, R. J.; Herrin, B. R.; Acharya, S.; Cooper, M. D.; Wilson, I. A., VLR Recognition of TLR5 Expands the Molecular Characterization of Protein Antigen Binding by Non-Ig-based Antibodies. *J Mol Biol* **2018,** *430* (9), 1350-1367.
25. Langefeld, T.; Mohamed, W.; Ghai, R.; Chakraborty, T., Toll-like receptors and NOD-like receptors: domain architecture and cellular signalling. *Adv Exp Med Biol* **2009,** *653*, 48-57.
26. Rosenstiel, P.; Jacobs, G.; Till, A.; Schreiber, S., NOD-like receptors: ancient sentinels of the innate immune system. *Cell Mol Life Sci* **2008,** *65* (9), 1361–1377.
27. Proell, M.; Riedl, S. J.; Fritz, J. H.; Rojas, A. M.; Schwarzenbacher, R., The Nod-like receptor (NLR) family: a tale of similarities and differences. *PloS One* **2008,** *3* (4), e2119.
28. Kaparakis, M.; Philpott, D. J.; Ferrero, R. L., Mammalian NLR proteins; discriminating foe from friend. *Immunol Cell Biol* **2007,** *85* (6), 495–502.
29. Kufer, T. A.; Banks, D. J.; Philpott, D. J., Innate immune sensing of microbes by Nod proteins. *Ann N Y Acad Sci* **2006,** *1072*, 19–27.
30. Ausubel, F. M., Are innate immune signaling pathways in plants and animals conserved? *Nat Immunol* **2005,** *6* (10), 973–979.
31. Nurnberger, T.; Brunner, F.; Kemmerling, B.; Piater, L., Innate immunity in plants and animals: striking similarities and obvious differences. *Immunol Rev* **2004,** *198*, 249-266.
32. Janeway, C. A., Jr.; Medzhitov, R., Innate immune recognition. *Annu Rev Immunol* **2002,** *20*, 197–216.
33. Nurnberger, T.; Brunner, F., Innate immunity in plants and animals: emerging parallels between the recognition of general elicitors and pathogen-associated molecular patterns. *Curr Opin Plant Biol* **2002,** *5* (4), 318–324.
34. Hayashi, F.; Smith, K. D.; Ozinsky, A.; Hawn, T. R.; Yi, E. C.; Goodlett, D. R.; Eng, J. K.; Akira, S.; Underhill, D. M.; Aderem, A., The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. *Nature* **2001,** *410* (6832), 1099–1103.
35. Kajander, T.; Sachs, J. N.; Goldman, A.; Regan, L., Electrostatic interactions of Hsp-organizing protein tetratricopeptide domains with Hsp70 and Hsp90: computational analysis and protein engineering. *J Biol Chem* **2009,** *284* (37), 25364-25374.

36. Cortajarena, A. L.; Yi, F.; Regan, L., Designed TPR modules as novel anticancer agents. *ACS Chem Biol* **2008,** *3* (3), 161–166.
37. Jernigan, K. K.; Bordenstein, S. R., Tandem-repeat protein domains across the tree of life. *PeerJ* **2015,** *3*, e732.
38. Allan, R. K.; Ratajczak, T., Versatile TPR domains accommodate different modes of target protein recognition and function. *Cell Stress Chaperones* **2011,** *16* (4), 353-367.
39. Nyarko, A.; Mosbahi, K.; Rowe, A. J.; Leech, A.; Boter, M.; Shirasu, K.; Kleanthous, C., TPR-Mediated self-association of plant SGT1. *Biochemistry* **2007,** *46* (40), 11331-11341.
40. Magliery, T. J.; Regan, L., Beyond consensus: statistical free energies reveal hidden interactions in the design of a TPR motif. *J Mol Biol* **2004,** *343* (3), 731-745.
41. Main, E. R.; Xiong, Y.; Cocco, M. J.; D'Andrea, L.; Regan, L., Design of stable alpha-helical arrays from an idealized TPR motif. *Structure* **2003,** *11* (5), 497–508.
42. Xing, H.; Fu, X.; Yang, C.; Tang, X.; Guo, L.; Li, C.; Xu, C.; Luo, K., Genome-wide investigation of pentatricopeptide repeat gene family in poplar and their expression analysis in response to biotic and abiotic stresses. *Sci Rep* **2018,** *8* (1), 2817.
43. Shen, C.; Wang, X.; Liu, Y.; Li, Q.; Yang, Z.; Yan, N.; Zou, T.; Yin, P., Specific RNA Recognition by Designer Pentatricopeptide Repeat Protein. *Mol Plant* **2015,** *8* (4), 667-670.
44. Manna, S., An overview of pentatricopeptide repeat proteins and their applications. *Biochimie* **2015,** *113*, 93-99.
45. Stothard, P., The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques* **2000,** *28* (6), 1102-1104.
46. Holowachuk, E. W.; Ruhoff, M. S., Efficient gene synthesis by Klenow assembly/extension-Pfu polymerase amplification (KAPPA) of overlapping oligonucleotides. *PCR methods and applications* **1995,** *4* (5), 299-302.
47. Gasteiger, E.; Hoogland, C.; Gattiker, A.; Duvaud, S. e.; Wilkins, M. R.; Appel, R. D.; Bairoch, A., Protein Identification and Analysis Tools on the ExPASy Server. In *The Proteomics Protocols Handbook*, Walker, J. M., Ed. Humana Press: Totowa, NJ, 2005; pp 571-607.
48. Crooks, G. E.; Hon, G.; Chandonia, J. M.; Brenner, S. E., WebLogo: a sequence logo generator. *Genome Res* **2004,** *14* (6), 1188–1190.
49. Sullivan, B. J.; Durani, V.; Magliery, T. J., Triosephosphate Isomerase by Consensus Design: Dramatic Differences in Physical Properties and Activity of Related Variants. *J Mol Biol* **2011,** *413* (1), 195-208.
50. Magliery, T. J.; Regan, L., Sequence variation in ligand binding sites in proteins. *BMC Bioinformatics* **2005,** *6*, 240.
51. Lee, S.-C.; Park, K.; Han, J.; Lee, J.-j.; Kim, H. J.; Hong, S.; Heu, W.; Kim, Y. J.; Ha, J.-S.; Lee, S.-G.; Cheong, H.-K.; Jeon, Y. H.; Kim, D.; Kim, H.-S., Design of a binding scaffold based on variable lymphocyte receptors of jawless vertebrates by module engineering. *Proceedings of the National Academy of Sciences* **2012,** *109* (9), 3299.
52. Skerra, A., Engineered protein scaffolds for molecular recognition. *J Mol Recognit* **2000,** *13* (4), 167–187.
53. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J., Basic local alignment search tool. *J Mol Biol* **1990,** *215* (3), 403–10.
54. Boch, J.; Scholze, H.; Schornack, S.; Landgraf, A.; Hahn, S.; Kay, S.; Lahaye, T.; Nickstadt, A.; Bonas, U., Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors. *Science* **2009,** *326* (5959), 1509.

4.7 Supplemental Information



SI Figure 4.1. Mutual information for *Hydra magnipapillata* 42PRs.



SI Figure 4.2. Calibration curve from known standards for Superdex S75 analytical column.

Chapter 5. Consensus Design of a Family 1 Cellulose Binding Module

(In preparation for publication)

Jennifer P. McCord, V. Grey Fritz, Jianzhao Liu, Alan R. Esker, Tijana Z. Grove

Department of Chemistry, Virginia Tech, Blacksburg, VA 24061

Keywords: cellulose binding module, protein design, protein engineering, biofuel, quartz crystal microbalance

5.1 Abstract

Lignocellulose is the most abundant and inexpensive source of biomass and therefore is widely considered a possible source for liquid fuel. However, processing costs have kept lignocellulosic fuels from competing commercially with starch-based biofuels. In recent years a strategy to protect processing enzymes with synergistic proteins emerged to reduce the amount of enzyme necessary for lignocellulosic biofuel production. Simultaneously, protein engineering approaches have emerged to optimize proteins for function and stability, enabling the use of proteins under non-native conditions required for any necessary application. Here we present the consensus design of a protein based on the carbohydrate-binding protein domain CBM1 that binds to cellulosic materials. The resulting designed protein is a stable monomeric protein that binds to both microcrystalline cellulose and amorphous regenerated cellulose thin films. By studying small changes to the binding site, we can better understand how these proteins bind to different cellulose-based materials in nature, and how to apply their use to industrial applications, such as enhancing the saccharification of lignocellulosic feedstock for biofuel production.

5.2 Introduction

Plant-based liquid fuels are valuable renewable replacements for fossil fuels, with the potential to reduce societal dependence on petroleum.[1] Most commonly, they are made from starchy plants such as corn, which are energetically expensive and water-hungry plants to grow. Cellulosic feedstocks in contrast are abundant, inexpensive and bio-renewable; however, the processing required to convert these feedstocks into usable simple sugars is currently energetically expensive and thus keeps the price of cellulosic ethanol relatively high.

The cellulases that convert lignocellulose to glucose are the most expensive part of lignocellulosic ethanol, still accounting for roughly 15% of the cost per gallon even after a decade of cost-reduction strategies by enzyme production companies.[2] To keep this cost as low as possible, pretreatment to improve the accessibility of the cellulose is necessary. However, pretreatment produces a variety of compounds that inhibit cellulase efficiency, hindering efficient saccharification.[3] In recent years, a strategy that relies on synergistic (helper) proteins to protect cellulases has emerged to reduce the amount of enzyme necessary for pretreatment.[4-6] The mechanism of action of these synergistic proteins is not well-understood, but is believed to rely on increasing the accessibility of cellulose to the cellulases through modification of the crystalline cellulose structure.[5] We hypothesize that improving the binding of the additives to cellulose will improve their activity.

Some cellulases contain domains responsible for specific binding of the enzyme onto cellulose. These domains, cellulose binding modules (CBMs), are found in many cellulose-active enzymes and are responsible for increasing the surface concentration of enzyme on cellulose, thereby enhancing enzymatic activity. To date, the CAzy database contains 84 different families

of CBMs. Of the CBMs, those from family 1 are found almost exclusively in fungi and are the smallest at 36 amino acids long. They consist of three beta strands with two disulfide bonds, resulting in a wedge-shaped structure. CBM1 proteins and peptides have been proposed for a variety of uses, including additives for lignocellulose pretreatment,[7] or attachment to other cellulases to increase binding affinity to cellulose.[8] We set out to design a consensus CBM1 protein appropriate for these applications, which could be produced on a large scale.

This work describes the consensus design to create a small cellulose-binding protein based on carbohydrate-binding modules from family 1 that is stable, can be expressed in *E. coli*, has a well-defined tertiary structure, and retains its cellulose-binding function. The protein was expressed in the *E. coli* BL21 cells and purified using previously described ELP-intein purification system. This strategy minimizes hands-on purification time and materials, and allows for easily scaled up production. We studied four variants of designed protein, changing only two positions on the binding surface as indicated by sequence analysis, and found this had large effects on proteins binding and folding. To our knowledge, this is the first report of statistical design of a family 1 cellulose binding module.

5.3 Experimental Methods

5.3.1 Multiple Sequence Alignment (MSA) and Consensus Design

For our multiple sequence alignment, we used 1539 unique sequences for CBM1 from the CAZy[9] and SMART[10] databases. The sequences were aligned manually in Microsoft Excel, and counting functions were used to determine the occurrence of each amino acid at each position (**Figure 5.1a**). For our design, we chose to use the amino acid with the highest relative entropy (**Equation 5.1**) of each position rather than the most common amino acid, as this

approach has led to several particularly stable engineered proteins.[11] The relative entropy allows us to distinguish between residues important for stability and residues chosen for other factors. In this equation, D is the relative entropy, $p(x_i)$ is the probability of finding amino acid x at position i, and $q(x)$ is the probability of finding amino acid x in the yeast proteome.

$$D = p(x_i) \ln \frac{p(x_i)}{q(x)} \quad \textbf{(5.1)}$$

5.3.2 Mutual Information Analysis: Correlations Between Positions

We used a mutual information (MI) analysis to determine potential correlations between positions. MI is the relative entropy between the actual probability distribution for two positions and the independent probability distribution between those same positions. The MI between two sites i and j is **Equation 5.2**, where $p(x_i)$ is the probability of observing amino acid x at position i, $p(y_j)$ is the probability of observing amino acid y at position j, and $p(x_i, y_j)$ is the probability of observing amino acid x at position i and amino acid y at position j.

$$MI_{ij} = \sum_i \sum_j p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i) p(y_j)} \quad \textbf{(5.2)}$$

To determine the statistical significance of the MI values, the amino acids in each position were scrambled, and the MI of this scrambled dataset was calculated. The highest value from this scrambled data was taken as the noise.

5.3.3 Cloning

Designed protein sequences were translated to corresponding DNA sequences using the Bioinformatics Sequence Manipulation Suite and optimized for bacterial codon usage. Such

synthetic genes were synthesized by GENEWIZ (South Plainfield, NJ) and cloned into plasmid pET-ELP, purchased from Addgene, by ligation of restriction sites BsrGI and HindIII. Gene identity was confirmed by Sanger sequencing at the Biocomplexity Institute of Virginia Tech.

5.3.4 Protein Expression

Overnight cultures of BL-21 (DE3) cells were diluted 2:100 in 1 L of Luria broth media at 37°C, with shaking at 200 rpm, and were grown to an OD600 of 0.5–0.8. Expression was induced with 0.25 m*M* IPTG, followed by 16-20 h of expression at 17 °C. The cells were harvested by centrifugation at 5,000 rpm for 15 min and the pellets were frozen at −80 °C until purification. To purify proteins, the cell pellet was resuspended in lysis buffer consisting of 10 m*M* Tris, 2 mM EDTA, 0.1 mg/mL lysozyme, and 0.1% Tween 20, pH 8.5. After 40 s sonication at 30% power using a microtip and Mison sonifier, lysed cells were incubated on ice with 2 µL DNAse I for 30 min, then centrifuged at 16,000 rpm for 30 min and the protein supernatant was collected.

5.3.5 Protein Purification

Proteins were purified as reported previously for the ELP-intein system.[12] Briefly, an equal volume of 3 M sodium chloride was mixed with the protein supernatant. The resulting sample was vortexed and incubated at 37 °C for 20 min, followed by centrifugation at 14,000 x g at 37°C for 10 min. The resulting pellet was resuspended in intein cleavage buffer consisting of PBS with 40 mM Bis-Tris, 2 mM EDTA, pH 6.2. This suspension was incubated at room temperature for 16-20 h, then mixed with an equal volume of 3 M sodium chloride, heated to 37°C for 20 min, and centrifuged at 14,000 x g at 37 °C for 10 min. Protein identity was confirmed with Matrix-Assisted Laser Desorption Ionization- Time of Flight Mass Spectrometry

(MALDI-TOF) indicating a molecular weight of 4055 Da, for a calculated molecular weight of 4038 Da, indicating the addition of a hydroxide group. Proteins were quantified by absorption at 280 nm using extinction coefficients calculated from the amino acid sequence using the Expasy Protparam tool. Proteins were dialyzed into NanoPure water using Spectra/Por 6 dialysis membrane tubing with a molecular weight cutoff of 1kDa and lyophilized before further use. After purification and dialysis, a protein yield of 5 mg/L was obtained.

5.3.6 Size Exclusion Chromatography

An Akta Prime Plus FPLC was used for size exclusion chromatography. Characterization of protein purity was done on the Superdex 75 16/600 Prep Grade column in 10 mM sodium chloride and 10 m*M* sodium phosphate buffer pH 7.4 at a flow rate of 1 ml/min.

5.3.7 Circular Dichroism

CD spectra were acquired using 5–10 µ*M* protein samples in NanoPure water using a Jasco J-815 CD spectrometer. Far-UV CD (190–260 nm) spectra were recorded at 25 °C to assess secondary structure. Each sample was recorded three times, from 190 to 260 nm in a 2 mm pathlength cuvette, and averaged. Data collected using a 1 nm bandwidth, 2 nm data pitch, and a data integration time of 1 s, was normalized to units of mean residue ellipticity for all samples. Thermal denaturation curves were recorded by monitoring molar ellipticity at 224 nm while heating from 30 to 90 °C in 2 °C increments with an equilibration time of 5 min at each temperature.

5.3.8 Binding Affinity Characterization

5.3.8.1 Binding to Avicel Microcellulose as a Model for Crystalline Cellulose

All binding experiments were performed in triplicate; reported error bars are one standard deviation. Samples of protein (typically 5-100 μM) were mixed in 1.5 mL Eppendorf tubes with an aqueous suspension of Avicel microcrystalline cellulose to a total concentration of 1 mg/mL and volume of 1 mL. Tubes were vortexed, then mixed by rotation at 4 °C for 20 h. Tubes were centrifuged at 10,000 RPM at 4 °C for 10 minutes to sediment the cellulose. The supernatant with unbound protein was collected and protein concentration was measured at 280 nm, using extinction coefficients calculated from the protein sequence by the Expasy Protparam tool.[13] The data was fit using nonlinear regression analysis with Origin software using the Langmuir binding model (**Equation 5.3**). In this equation, B is the measured amount of bound protein, B_{max} is the maximum theoretical amount of bound protein, F is the concentration of free protein, and K_d is the dissociation constant.

$$B = \frac{B_{max} F}{K_d + F} \qquad \textbf{(5.3)}$$

5.3.8.2 Binding to Regenerated Cellulose Thin Film as a Model for Amorphous Cellulose

Gold coated quartz crystal chips were cleaned by immersing into a mixture of ammonium hydroxide, hydrogen peroxide, and water (1:1:5, volume ratio) at 80 ^{o}C for 1 h, rinsed with water and dried with nitrogen gas. Trimethylsilyl cellulose (TMSC) was dissolved in toluene at 1 wt%, and was filtered through a 0.45 μM syringe filter (VWR PTFE), to obtain a transparent TMSC solution. The TMSC film was prepared by spin-coating the TMSC solution on a clean gold coated quartz crystal chip at 4000 rpm for 1 min, followed by vacuum-drying at room temperature overnight. The conversion of TMSC thin films to regenerated cellulose thin films for QCM-D study was achieved by exposing TMSC thin films to 10 wt% HCl vapor for 3min.

A regenerated cellulose film in the QCM-D cell was initially equilibrated with NanoPure water at a flow rate of 0.1 mL/min to obtain a stable baseline. Then, 6.2 μM rCBM1-YW in NanoPure water was pumped into the cell at 0.1 mL/min for 15 min, and the flow was paused for 20 min to allow the rCBM1 in the cell to establish a dynamic equilibrium onto the regenerated cellulose surface. In order to remove the reversibly adsorbed rCBM1-YW, the NanoPure water was introduced into the cell for ~30 min. Once a new baseline was achieved after the NanoPure water flowed through the cell, an rCBM1-YW solution with the next higher concentration was flowed over the regenerated cellulose surface. This process was repeated in succession from lowest to highest concentration (6.2-91.0 μM). The data was again fit using nonlinear regression analysis with Origin software using the Langmuir binding model (**Equation 5.3**).

5.4 Results and Discussion

5.4.1 Multiple Sequence Alignment (MSA) and Consensus Design

Consensus design is an efficient way to design stable proteins while retaining the natural function of the protein.[14-15] The consensus sequence of all CBM1s was determined primarily from the amino acid with the highest relative entropy at each position of the alignment of 1285 unique sequences. Previous experimental studies have found the aromatic residues at positions 5, 31, and 32, along with N29 and Q34, are primarily responsible for the specific interaction with cellulose.[16-22] This is consistent with our statistical analysis, which shows high levels of conservation for each of these positions. MSA analysis (**Figure 5.1a**) showed that positions 5 and 31 had roughly equivalent incidences of tryptophan and tyrosine. We used a mutual information analysis[11] (**Figure 5.1b**) to determine potential correlations between these positions. Such correlations often indicate positions necessary for packing in the protein core.[23] Since these positions are instead expected to be on the binding surface of our protein, combinations of

residues in these positions that occur more often than chance could indicate that these combinations were related to the cellulosic structures or materials each CBM in our MSA bound. Our analysis, however, indicated that these positions were not correlated with one another; that is, combinations of amino acids in these positions occurred approximately as chance would suggest. As a result, we decided to make four total sequences, with tryptophan and tyrosine each at positions 5 and 31 (**Figure 5.1c**). The final sequences are found in **Table 5.1**, with the tryptophan and tyrosine residues changing in bold.

Table 5.1. rCBM1 protein sequences.

Protein Name	Sequence
rCBM1-WW	QAAA**W**GQCGGIGWTGPTTCASGYTCTVQND**W**YSQCL
rCBM1-WY	QAAA**Y**GQCGGIGWTGPTTCASGYTCTVQND**W**YSQCL
rCBM1-YW	QAAA**W**GQCGGIGWTGPTTCASGYTCTVQND**Y**YSQCL
rCBM1-YY	QAAA**Y**GQCGGIGWTGPTTCASGYTCTVQND**Y**YSQCL

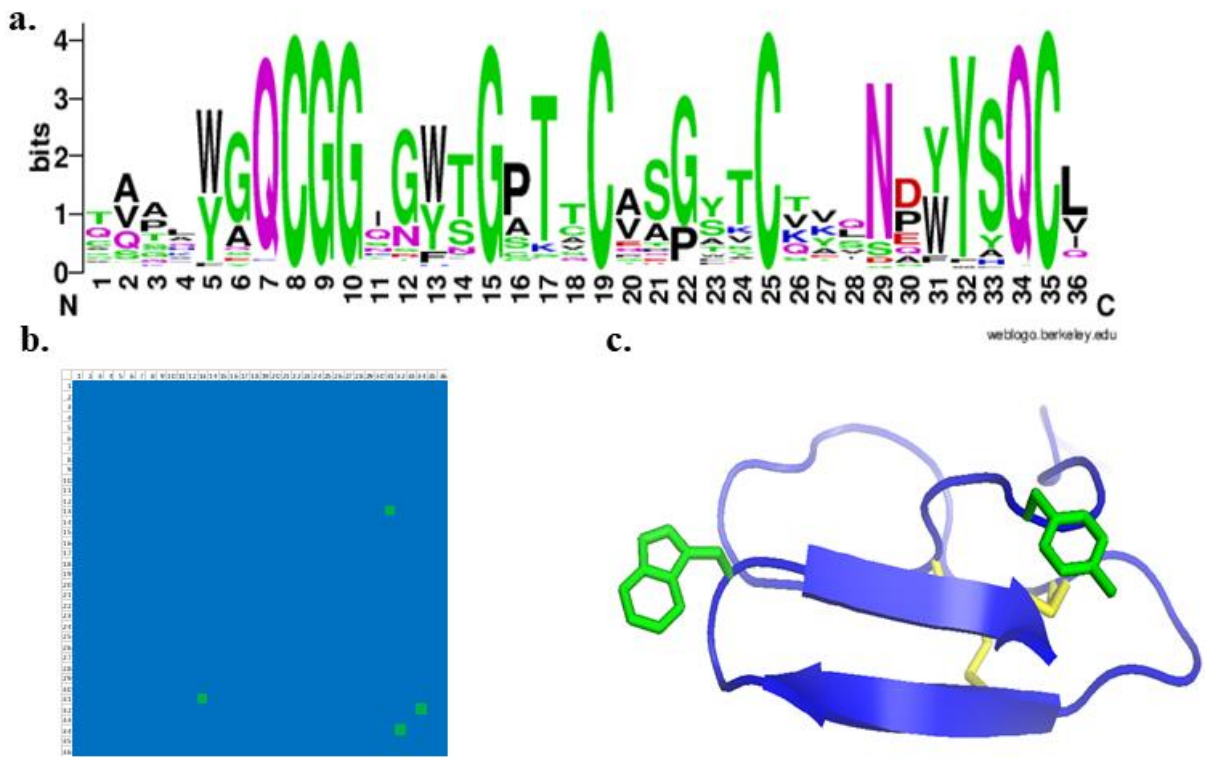


Figure 5.1. a. Multiple sequence alignment of CBM1. The height of individual one-letter amino acid codes indicates the frequency of occurrence of residues at each position. b. Mutual information for the CBM1 MSA indicates the probability relative to random chance that two amino acids occur at positions simultaneously.[11] Blue indicates mutual information values below the value of the noise; green indicates values between the noise and one standard deviation above the noise. c. MUSTER model of whole protein rCBM1-YW.[24]

5.4.2 Protein Purification and Characterization

rCBM1 proteins were expressed at 17 °C in BL-21 cells. We chose to use the ELP-intein protein purification system developed by the Wood lab[12, 25] for its ability to scale up protein production. Because it is chromatography-free, large quantities of protein can be obtained using this system from a single round of purification. After ELP-intein purification, we assessed purity by size-exclusion chromatography. Each rCBM1 protein elutes in one peak with a small peak

also appearing at the void volume (**SI Figure 5.1**), which is likely residual ELP tag. MALDI indicated a molecular mass of 4055 Da, in good agreement with the expected molecular weight.

5.4.3 Secondary Structure Analysis

CD Spectroscopy was used to measure secondary structure content. For theoretical CD spectra, we used the DiChroCalc computational tool[26] to calculate spectra based on the Multi-Sources ThreadER (MUSTER)[24] model of the protein sequences (**Figure 5.2a**), which matched well with CBM1 crystal structures available in the PDB.[27] As expected from the crystal structures, the theoretical CD spectra showed a characteristic dip around 220, corresponding to a primarily β-sheet structure.

The experimental CD results were more surprising. Khazanov et al. previously characterized a recombinant CBM1 from *T. reesei* cellobiohydrolase I by CD, and found that it has a primarily random coil structure, which, as in our case, differed from their primarily β-sheet theoretical spectra.[28] We found by contrast that the experimental CD spectra of each of our designed protein indicates an alpha helical fold, with typical peaks at 208 and 220 nm (**Figure 5.2b**). While all proteins had roughly the same overall structure, rCBM1-YW showed much stronger evidence of secondary structure than the other proteins. As only the binding surface was changing, we did not expect the two point mutations to have a large effect on protein structure.

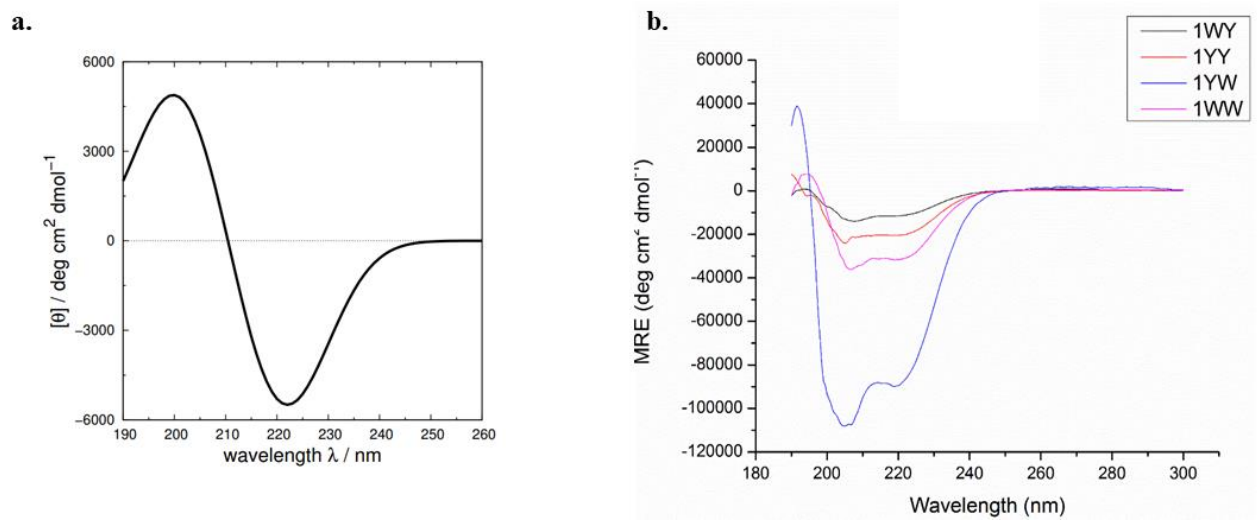


Figure 5.2. rCBM1 CD Characterization. a. Theoretical CD Spectra of CBM1. b. Experimental CD spectra, in which the pink trace is rCBM1-WW, the black trace is rCBM1-WY, the blue trace is rCBM1-YW, and the red trace is rCBM1-YY.

Our first hypothesis was that improper disulfide bond formation may be the driving force behind this unusual secondary structure formation, and that this may affect binding through burial of the aromatic residues in the protein interior. SEC and MALDI data appear to rule out intermolecular disulfide bond formation, as together they suggest a single, monomeric protein species in solution. Fluorescence spectroscopy under native and reducing conditions (**SI Figure 5.2**) indicates no significant change in environment for the aromatic residues. This indicates that the aromatic residues are indeed on the protein surface, and that the disulfide bond formation does not significantly impact these residues, which are important for binding. We hypothesize that post-translational glycosylation in natural CBM1s on which the theoretical CD spectra are based may have the effect of stabilizing the primarily β-sheet structures found in nature. As the residues important for binding remained on the surface, we moved on to characterize the affinity of rCBM1-YW to cellulose.

5.4.4 Binding Affinity to Avicel and Regenerated Cellulose Thin Films

Previous work has shown that the CBM1 family of proteins bind to a variety of cellulosic materials and structures, with CBMs from different enzymes have preferential affinities for different cellulose structures and materials.[16, 29-31] Guo and Catchmark demonstrated the preference of the CBM from *Tichoderma Reesei* Cel_{6A} for the reducing ends of the cellulose, while the CBM from Cel_{7A} showed no such preference in the same study.[16]

As a model for crystalline cellulose, we used commercially available Avicel microcellulose without further purification. The affinities of all rCBM1 proteins to Avicel were assessed with adsorption isotherm measurements using the Langmiur model (**Figure 5.3a**). rCBM1-YW had a K_d= 45±17μM to Avicel, within error of what has been published for *T. Reesei* Cel_{6A} and Cel_{7A}.[16] Interestingly, we did not observe binding to the other proteins (**SI Figure 5.1**). We hypothesize that the lack of secondary structure in the other variants is responsible for this lack of binding.

rCBM1-YW binding to an amorphous cellulose thin film was assessed in triplicate using quartz crystal microbalance with dissipation monitoring (QCM-D). The adsorption behavior was monitored *in situ* and in real time. rCBM1-YW was loaded with concentrations from 6.2-100 μM, allowed to equilibrate, rinsed with NanoPure water and allowed to equilibrate a second time. The mass of the rCBM1-YW remaining on the thin film at his second equilibrium was used to build the adsorption isotherm (**Figure 5.3b**).

The change in mass per unit area (Δm) for QCM-D measurements can be calculated from the Sauerbrey equation (**Equation 5.4**), where Δm is the change in mass per unit area on the

sensor corresponding to a measured change in frequency (Δf) in this equation, *n* is the overtone number and *C* is a constant (0.177 mg * m^{-2} * Hz^{-1}).[32]

$$\Delta m = -\frac{C\Delta f}{n} \quad \textbf{(5.4)}$$

The calculation of the mass of protein on the cellulose surface assumes all cellulose was involved in binding. We used this model for the change in mass to construct an adsorption isotherm using the 5th overtone as published previously.[33] Fitting to the Langmuir binding model resulted in a K_d = 19 ± 4 µM, which is again within error to what was measured for the wild type.[31]

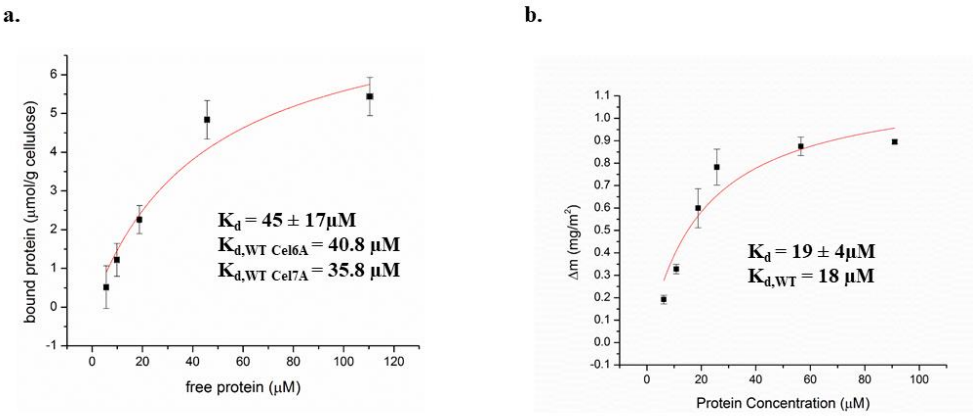


Figure 5.3. rCBM1-YW binding to a. Avicel microcrystalline cellulose by adsorption isotherm measurements and b. amorphous regenerated cellulose by quartz crystal microbalance.

There have been several efforts in the past few years to design small peptide mimics of CBMs, rationally and by phage display.[28, 34-35] These bind to cellulose quite well (K_d ~ 10^{-5} M), but lack the tertiary structure that contributes to the binding affinity and stability of larger CBMs.[35] While their smaller size is desirable for some applications, the better stability and

chromatography-free purification of rCBM1 are more convenient for applications requiring large-scale production.

5.5 Conclusions

In this work, we describe the design and characterization of a consensus cellulose binding module from family 1 to better understand and improve upon natural cellulose binding modules. We propose that designed CBM1s could improve upon currently proposed synergistic additives to the cellulase cocktail used in enzyme saccharification of lignocellulosic biomass, or be used as a possible domain in designer cellulases. Modifying the surface residues that interact directly with cellulose had a surprisingly large effect on the structure and binding to cellulose. rCBM1-YW had a binding affinity to microcrystalline and amorphous cellulose that matched that of naturally occurring CBM1s, while increasing thermal and chemical stability.

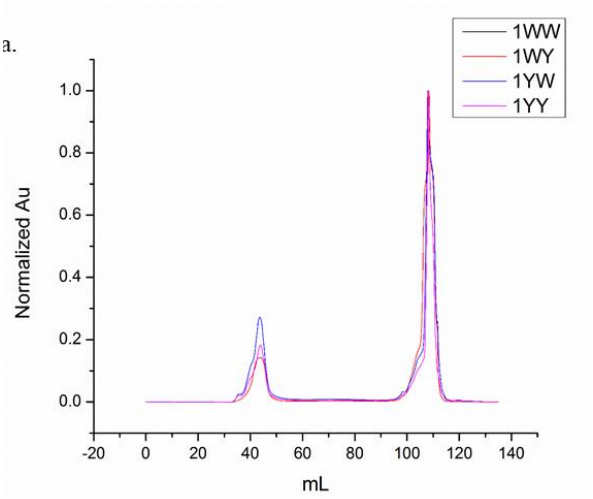
5.7 References

1. Somma, D.; Lobkowicz, H.; Deason, J. P., Growing America's fuel: an analysis of corn and cellulosic ethanol feasibility in the United States. *Clean Techn Environ Policy* **2010,** *12*.
2. Várnai, A.; Siika-aho, M.; Viikari, L., Carbohydrate-binding modules (CBMs) revisited: reduced amount of water counterbalances the need for CBMs. *Biotechnology for Biofuels* **2013,** *6* (1), 30.
3. Várnai, A.; Siika-aho, M.; Viikari, L., Restriction of enzymatic hydrolysis of pretreated spruce by lignin. *Enzyme Microb Technol* **2010,** *46*.
4. Oliveira, C.; Romaní, A.; Gomes, D.; Cunha, J. T.; Gama, F. M.; Domingues, L., Recombinant family 3 carbohydrate-binding module as a new additive for enhanced enzymatic saccharification of whole slurry from autohydrolyzed Eucalyptus globulus wood. *Cellulose* **2018,** *25* (4), 2505-2514.
5. Kim, I. J.; Lee, H. J.; Choi, I.-G.; Kim, K. H., Synergistic proteins for the enhanced enzymatic hydrolysis of cellulose by cellulase. *Applied Microbiology and Biotechnology* **2014,** *98* (20), 8469-8480.
6. Cunha, E. S.; Hatem, C. L.; Barrick, D., Synergistic enhancement of cellulase pairs linked by consensus ankyrin repeats: Determination of the roles of spacing, orientation, and enzyme identity. *Proteins* **2016,** *84* (8), 1043-1054.
7. Mello, B. L.; Polikarpov, I., Family 1 carbohydrate binding-modules enhance saccharification rates. *AMB Express* **2014,** *4*, 36-36.
8. Shoseyov, O.; Shani, Z.; Levy, I., Carbohydrate Binding Modules: Biochemical Properties and Novel Applications. *Microbiology and Molecular Biology Reviews* **2006,** *70* (2), 283-95.
9. The Carbohydrate-Active enZYmes Database. : ; http://www.cazy.org.
10. Schultz, J.; Milpetz, F.; Bork, P.; Ponting, C. P., SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A* **1998,** *95* (11), 5857–64.
11. Durani, V.; Magliery, T. J., Protein engineering and stabilization from sequence statistics: variation and covariation analysis. *Methods Enzymol* **2013,** *523*, 237–56.
12. Wu, W.-Y.; Fong, B. A.; Gilles, A. G.; Wood, D. W., Recombinant Protein Purification by Self-Cleaving Elastin-like Polypeptide Fusion Tag. *Current Protocols in Protein Science* **2009,** *58* (1), 26.4.1-26.4.18.
13. Gasteiger, E.; Hoogland, C.; Gattiker, A.; Duvaud, S. e.; Wilkins, M. R.; Appel, R. D.; Bairoch, A., Protein Identification and Analysis Tools on the ExPASy Server. In *The Proteomics Protocols Handbook*, Walker, J. M., Ed. Humana Press: Totowa, NJ, 2005; pp 571-607.
14. Sullivan, B. J.; Durani, V.; Magliery, T. J., Triosephosphate Isomerase by Consensus Design: Dramatic Differences in Physical Properties and Activity of Related Variants. *J Mol Biol* **2011,** *413* (1), 195-208.
15. Parker, R.; Mercedes-Camacho, A.; Grove, T. Z., Consensus design of a NOD receptor leucine rich repeat domain with binding affinity for a muramyl dipeptide, a bacterial cell wall fragment. *Protein Sci* **2014,** *23* (6), 790–800.

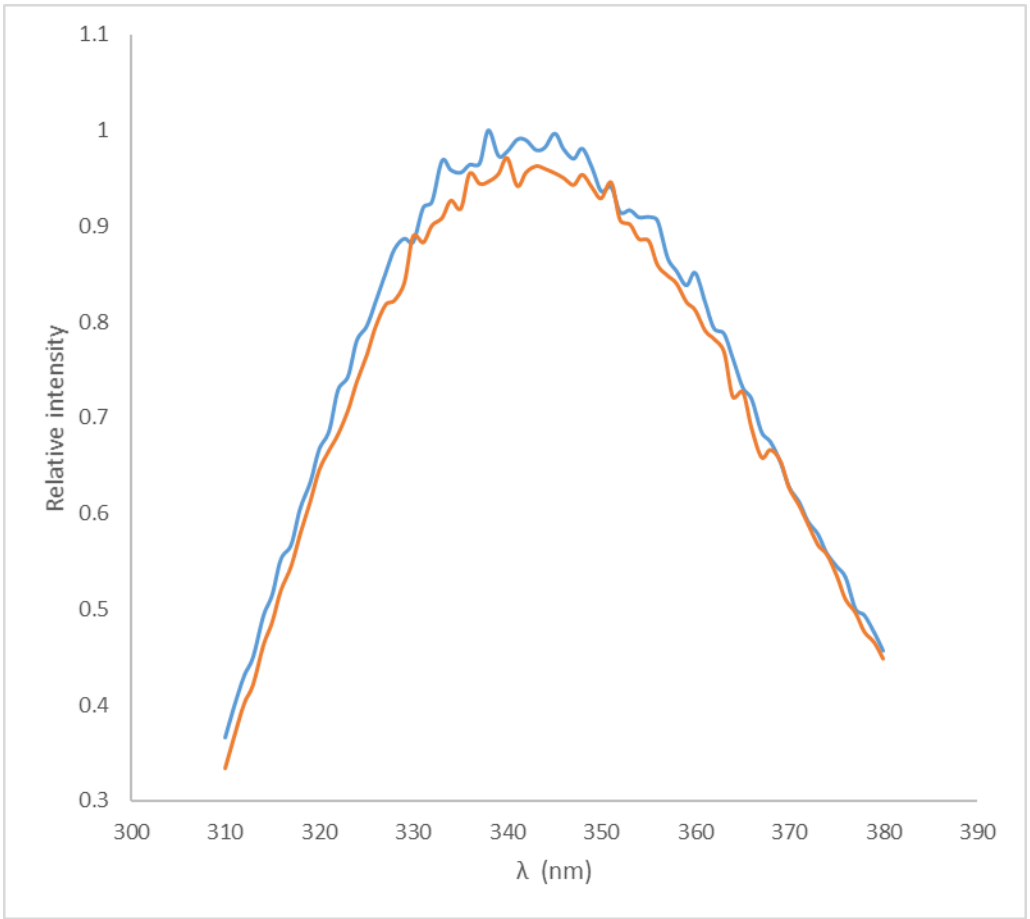
16. Guo, J.; Catchmark, J. M., Binding Specificity and Thermodynamics of Cellulose-Binding Modules from Trichoderma reesei Cel7A and Cel6A. *Biomacromolecules* **2013,** *14* (5), 1268-1277.
17. Sipos, B., Characterisation of specific activities and hydrolytic properties of cell-wall degrading enzymes produced by Trichoderma reesei Rut C30 on different carbon sources. *Appl Biochem Biotechnol* **2009,** *161*.
18. Palonen, H.; Tjerneld, F.; Zacchi, G.; Tenkanen, M., Adsorption of Trichoderma reesei CBH I and EG II and their catalytic domains on steam pretreated softwood and isolated lignin. *J Biotechnol* **2004,** *107*.
19. Suurnäkki, A., Trichoderma reesei cellulases and their core domains in the hydrolysis and modification of chemical pulp. *Cellulose* **2000,** *7*.
20. Linder, M.; Mattinen, M.-L.; Kontteli, M.; Lindeberg, G.; Ståhlberg, J.; Drakenberg, T.; Reinikainen, T.; Pettersson, G.; Annila, A., Identification of functionally important amino acids in the cellulose-binding domain of Trichoderma reesei cellobiohydrolase I. *Protein Science* **1995,** *4* (6), 1056-1064.
21. Brun, E.; Moriaud, F.; Gans, P.; Blackledge, M. J.; Barras, F.; Marion, D., Solution Structure of the Cellulose-Binding Domain of the Endoglucanase Z Secreted by Erwinia chrysanthemi. *Biochemistry* **1997,** *36* (51), 16074-16086.
22. Beckham, G. T.; Matthews, J. F.; Bomble, Y. J.; Bu, L.; Adney, W. S.; Himmel, M. E.; Nimlos, M. R.; Crowley, M. F., Identification of Amino Acids Responsible for Processivity in a Family 1 Carbohydrate-Binding Module from a Fungal Cellulase. *The Journal of Physical Chemistry B* **2010,** *114* (3), 1447-1453.
23. Sullivan, B. J.; Nguyen, T.; Durani, V.; Mathur, D.; Rojas, S.; Thomas, M.; Syu, T.; Magliery, T. J., Stabilizing Proteins from Sequence Statistics: The Interplay of Conservation and Correlation in Triosephosphate Isomerase Stability. *J Mol Biol* **2012,** *420* (4–5), 384–399.
24. Wu, S.; Zhang, Y., MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins* **2008,** *72* (2), 547-56.
25. Banki, M. R.; Feng, L.; Wood, D. W., Simple bioseparations using self-cleaving elastin-like polypeptide tags. *Nat Methods* **2005,** *2*.
26. Bulheller, B. M.; Hirst, J. D., DichroCalc--circular and linear dichroism online. *Bioinformatics (Oxford, England)* **2009,** *25* (4), 539-40.
27. Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M., The protein data bank: A computer-based archival file for macromolecular structures. *J Mol Biol* **1977,** *112* (3), 535-542.
28. Khazanov, N.; Iline-Vul, T.; Noy, E.; Goobes, G.; Senderowitz, H., Design of Compact Biomimetic Cellulose Binding Peptides as Carriers for Cellulose Catalytic Degradation. *The Journal of Physical Chemistry B* **2016,** *120* (2), 309-319.
29. Lehtiö, J.; Sugiyama, J.; Gustavsson, M.; Fransson, L.; Linder, M.; Teeri, T. T., The binding specificity and affinity determinants of family 1 and family 3 cellulose binding modules. *Proc Natl Acad Sci U S A* **2003,** *100* (2), 484-489.
30. Segato, F.; Damasio, A. R.; Gonçalves, T. A.; Murakami, M. T.; Squina, F. M.; Polizeli, M.; Mort, A. J.; Prade, R. A., Two structurally discrete GH7-cellobiohydrolases compete for the same cellulosic substrate fiber. *Biotechnol. Biofuels* **2012,** *5*.

31. Zhang, Y.; Yang, F.; Hu, F.; Song, J.; Wu, S.; Jin, Y., Binding preference of family 1 carbohydrate binding module on nanocrystalline cellulose and nanofibrillar cellulose films assessed by quartz crystal microbalance. *Cellulose* **2018,** *25* (6), 3327-3337.
32. Sauerbrey, G., Verwendung von Schwingquarzen zur Wägung dünner Schichten und zur Mikrowägung. *Zeitschrift für Physik* **1959,** *155* (2), 206-222.
33. Liu, Z.; Choi, H.; Gatenholm, P.; Esker, A. R., Quartz Crystal Microbalance with Dissipation Monitoring and Surface Plasmon Resonance Studies of Carboxymethyl Cellulose Adsorption onto Regenerated Cellulose Surfaces. *Langmuir* **2011,** *27* (14), 8718-8728.
34. Serizawa, T.; Iida, K.; Matsuno, H.; Kurita, K., Cellulose-binding Heptapeptides Identified by Phage Display Methods. *Chemistry Letters* **2007,** *36* (8), 988-989.
35. Guo, J.; Catchmark, J. M.; Mohamed, M. N. A.; Benesi, A. J.; Tien, M.; Kao, T.-h.; Watts, H. D.; Kubicki, J. D., Identification and Characterization of a Cellulose Binding Heptapeptide Revealed by Phage Display. *Biomacromolecules* **2013,** *14* (6), 1795-1805.

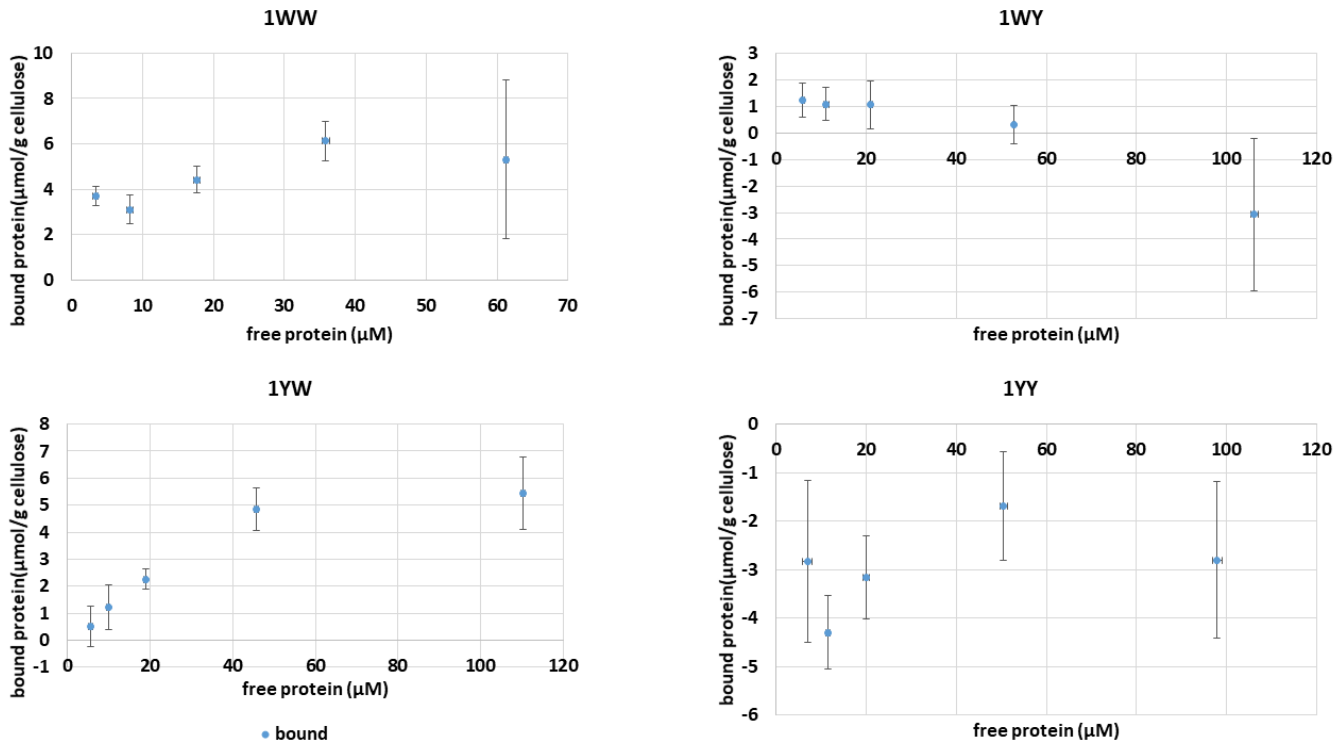
5.8 Supplemental Information



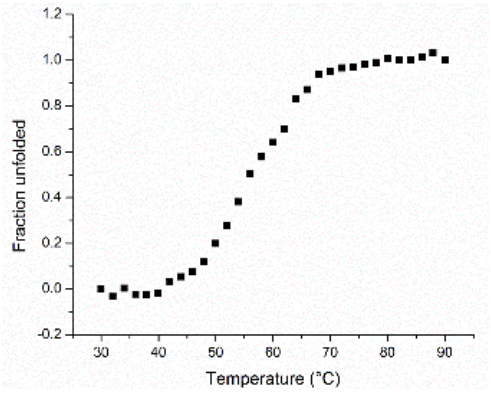
SI Figure 5.1. Size-exclusion chromatography traces for each protein.



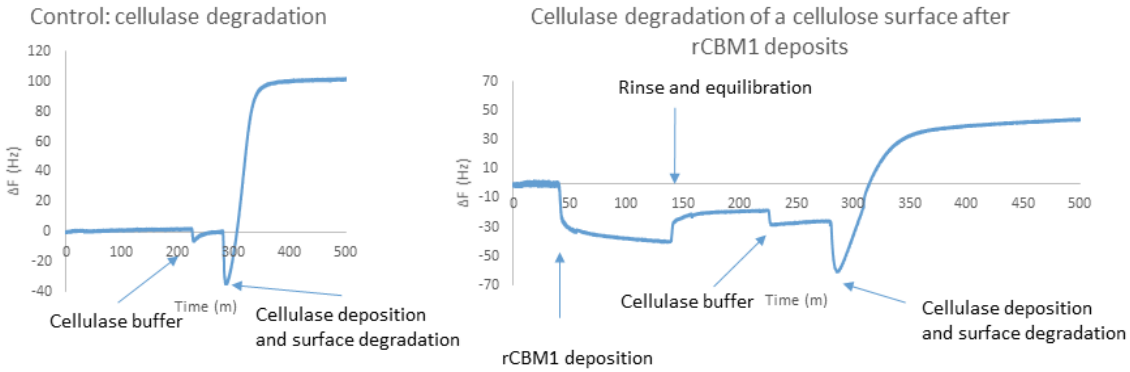
SI Figure 5.2. Fluorescence measurements of rCBM1-YW (blue trace) and rCBM1-YW in the presence of 5 mM dithiothreitol (red trace).



SI Figure 5.3. Adsorption isotherm measurements of all rCBM1 proteins.



SI Figure 5.4. The thermal denaturation of rCBM1-YW was followed using CD spectroscopy at 224 nm. This curve fitted to a two-state model shows a T_m of 57°C, approximately 7°C more stable than natural members of the CBM1 family.



SI Figure 5.6. QCM Competition Experiment. rCBM1-YW inhibits cellulase deposition on the cellulose surface.

Chapter 6. Engineering an Integrin Binding Site onto a Recombinant Keratin Scaffold

Jennifer P. McCord, Mark R. Van Dyke and Tijana Z. Grove

Department of Chemistry, Virginia Tech, Blacksburg, VA 24061

Keywords: keratin, biomaterials, integrin binding

6.1 Abstract

Biomaterials made from extracted keratin have mechanical and biochemical properties that make them ideal scaffolds for tissue engineering and wound healing. However, naturally extracted keratin materials contain unwanted byproducts. Recombinant keratin biomaterials are free from these disadvantages, while heterologous expression of these proteins allows us to manipulate the primary sequence. We endeavored to add an RGD sequence to facilitate cell adhesion to the recombinant keratin proteins.

6.2 Introduction

Biomaterials made from extracted keratin have mechanical and biochemical properties that make them ideal scaffolds for tissue engineering and wound healing.[1] The harsh conditions required to extract natural keratin leads to protein degradation, and yet byproducts such as melanin are difficult to remove. This lack of purity can lead to undesirable immune responses, and the inclusion of melanin makes the material unsuitable for potential ocular applications. Our lab has recently described the self-assembly of recombinant keratin proteins, which are free from these disadvantages.[2] Heterologous expression of these proteins additionally allows for the manipulation of the primary sequence of the protein. We describe an example sequence

modification to tune the properties of keratin biomaterials through integration of an integrin binding motif.

Cells primarily adhere to the extracellular matrix through attachment through the integrin receptors on the cell surface.[3] Of natural human hair keratin proteins, nearly 80% contain at least one integrin binding site.[4] This includes the K31 protein, which contains the LDV motif that binds to integrins $\alpha_4\beta_1$ and $\alpha_4\beta_7$.[3] However, the LDV motif is contained in the interior of the protein, limiting its ability to adhere to the receptors.[5] We endeavored to add an RGD sequence in a more accessible location to facilitate cell adhesion.

6.3 Materials and Methods

6.3.1 Gene design

Gene sequences corresponding to K31 and K81 in plasmid *pProExHtam*, which contains an N-terminal histidine affinity tag and an ampicillin resistance gene, were cloned previously.[2] The Megaprimer whole-plasmid cloning technique was used to add an RGD site to the K31 plasmid immediately after the BamHI restriction site.[6-7] A mutagenic primer with the desired insertion and an overlapping primer upstream of the desired mutation were used. After standard PCR with these primers and template plasmid using a high-fidelity polymerase, we ran a DNA agarose gel with ethidium bromide to confirm synthesis of the Megaprimer, and gel purified the product. The second round of PCR used the Megaprimer product of the first round of PCR along with more template plasmid. As with site-directed mutagenesis above, the template plasmid was then digested with DPN1. Gene sequence was confirmed by the Bioinformatics Institute of Virginia Tech (Blacksburg, VA.

6.3.2 Protein Expression and Purification

K31 and K81 were expressed in BL21 (DE3) *E. coli* cells. Cell cultures were grown for 16 h overnight in Luria Broth (LB) media at 37 °C with shaking at 250 rpm. Cells were then diluted 1:100 in LB media and grown to an optical density of 0.6–0.8 at which time protein expression was induced with 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG). Protein expression was conducted at 37 °C for 4 h. Following expression, cells were harvested through centrifugation at 5000 rpm for 15 min and the cell pellet was resuspended in lysis buffer, pH 8, containing 50 mM Tris HCl, 300 mM sodium chloride, and 1 % Tween 20, and then stored at −80 °C until purification. The desired protein was purified from inclusion bodies under denaturing conditions following a procedure adapted from Honda et al.[8] The resuspended cell pellet was first thawed in a 37 °C water bath. Following this step, 25 mg of lysozyme was added and the sample was incubated on ice for 30 min. Subsequently, 10 mM $MgCl_2$, 1 mM $MnCl_2$, and 10 μg mL^{-1} of DNase were added to the mixture and incubated on ice for 30 min. Following incubation, 25 mL of detergent buffer, pH 8, consisting of 20 mM Tris HCl, 200 mM NaCl, 1% Triton X-100, and 2 mM EDTA was added and mixed with the protein sample. The sample was then centrifuged at 5000 rpm for 15 min, and the supernatant removed. This step was repeated until a tight pellet of inclusion bodies was formed. After obtaining the desired inclusion body pellet, 30 mL of extraction buffer, pH 8, containing 10 mM Tris HCl, 2 mM EDTA, 8 M urea, 10 mM βME, and 1 protease inhibitor cocktail tablet was added to resuspend the pellet. The sample was then centrifuged at 16,000 rpm for 1 h. The supernatant was collected for purification using a standard Ni-NTA affinity purification protocol under denaturing and reducing conditions and eluted with 300 mM imidazole in lysis buffer with 8 M urea and 10 mM βME.

6.3.3 Gel Electrophoresis

Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) was used to estimate molecular weight of the purified protein. Samples were prepared in a 1:1 ratio of SDS buffer to protein and analyzed on a 10% acrylamide gel. A broad range protein marker from New England Biolabs (Ipswich, MA) was used as a standard that contains proteins from 200 to 10 kDa.

6.3.4 Dialysis

Following affinity purification and molecular weight verification by SDS-PAGE and MS analysis, K31 and K81 were individually dialyzed out of elution buffer, pH 8, containing 300 mM NaCl, 50 mM Tris HCl, 300 mM imidazole, 10 mM βME, and 8 M urea. In the first step of dialysis the protein was dialyzed against buffer, pH 8, with 10 mM Na_2HPO_4, 75 mM NaCl, 5 mM DTT, and 8 M urea. Four additional dialysis steps were completed with decreasing amounts of urea equal to 6, 4, 2, and 0 M. Each of the steps were completed at 2-h intervals except for the last step, which was allowed to equilibrate overnight. Following removal of urea, protein was dialyzed into 10 mM Na_2HPO_4 and 100 mM NaCl at pH 7.4 over an additional 20 hours before use.

6.4 Results and Discussion

Human hair keratins K31 and K81 are known to self-assemble into fibrous structures in nature. Previously, we demonstrated that recombinant human hair keratin proteins retained their ability to self-assemble into heterodimers.[2] This self-assembly is a necessary prerequisite to material formation.[9] In adding an integrin-binding

6.4.1 Addition of RGD to Keratin Primary Sequence

We chose to add the RGD sequence to the K31 protein due to its ability to homooligomerize into fibrous bundles. While the natural K31 sequence contains the LDV motif that binds to integrins $\alpha_4\beta_1$ and $\alpha_4\beta_7$,[3] the LDV motif is in the interior of the protein, likely contained within an alpha helix.[5] To mediate cell attachment, the integrin-binding site must be on the protein surface, arranged in a loop.[3]

6.4.2 Expression, Purification, and Solubilization of K31-RGD

K31-RGD expresses identically to unmodified K31, at 20 mg/L after purification. Ni-NTA purification results in reasonably pure protein (**Figure 6.1**). It is important that the addition of RGD to K31 not interfere with self-assembly of the K31-K81 heterodimer. Given the high cysteine content of these proteins this was not expected. Dialysis to remove denaturant and reducing agent indeed results in a soluble protein with and without its K81 pair, as previously reported for K31.[2, 5]

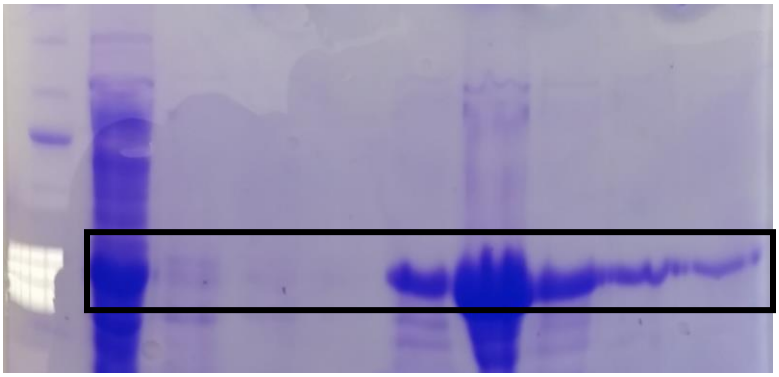


Figure 6.1. Ni-NTA purification of K31-RGD. The first lane is the ladder, followed by the flow-through and three washes, followed by five elutions with 300 mM imidazole.

6.5 Conclusions

Biomaterials made from extracted keratin have mechanical and biochemical properties that make them ideal scaffolds for tissue engineering and wound healing. Our lab has recently described the self-assembly of recombinant keratin proteins, which are free from these

disadvantages. Heterologous expression of these proteins allows for the manipulation of the primary sequence in order to tune the desired resulting biomaterial. We describe here an example sequence modification to tune the properties of keratin biomaterials through integration of an integrin binding motif. The addition of this motif does not affect the proteins' expression or ability to oligomerize as a homo- or heterodimer. We propose that doping recombinant keratin oligomers with modified properties into extracted keratin materials will result in materials with tunable properties for applications in tissue engineering, nerve regeneration, and wound healing.

6.6 References

1. Rouse, J. G.; Van Dyke, M. E., *A Review of Keratin-Based Biomaterials for Biomedical Applications*. Materials (Basel). 2010 Feb 3;3(2):999-1014. doi: 10.3390/ma3020999. eCollection 2010 Feb.: 2010.
2. Parker Rachael, N.; Roth Kristina, L.; Kim, C.; McCord Jennifer, P.; Van Dyke Mark, E.; Grove Tijana, Z., Homo- and heteropolymer self-assembly of recombinant trichocytic keratins. *Biopolymers* **2017,** *107* (10), e23037.
3. Ruoslahti, E., RGD AND OTHER RECOGNITION SEQUENCES FOR INTEGRINS. *Annual Review of Cell and Developmental Biology* **1996,** *12* (1), 697-715.
4. Sierpinski, P.; Garrett, J.; Ma, J.; Apel, P.; Klorig, D.; Smith, T.; Koman, L. A.; Atala, A.; Van Dyke, M., The use of keratin biomaterials derived from human hair for the promotion of rapid regeneration of peripheral nerves. *Biomaterials* **2008,** *29* (1), 118-128.
5. Basit, A.; asghar, F.; Sadaf, S.; Akhtar, M. W., Health improvement of human hair and their reshaping using recombinant keratin K31. *Biotechnology Reports* **2018,** *20*, e00288.
6. Tyagi, R.; Lai, R.; Duggleby, R. G., A new approach to 'megaprimer' polymerase chain reaction mutagenesis without an intermediate gel purification step. *BMC biotechnology* **2004,** *4*, 2.
7. Vander Kooi, C. W., Chapter Twenty One - Megaprimer Method for Mutagenesis of DNA. In *Methods Enzymol*, Lorsch, J., Ed. Academic Press: 2013; Vol. 529, pp 259-269.
8. Honda, Y.; Koike, K.; Kubo, Y.; Masuko, S.; Arakawa, Y.; Ando, S., In vitro Assembly Properties of Human Type I and II Hair Keratins. *Cell Structure and Function* **2014,** *39* (1), 31-43.
9. Stoppel, W. L.; Ghezzi, C. E.; McNamara, S. L.; III, L. D. B.; Kaplan, D. L., Clinical Applications of Naturally Derived Biopolymer-Based Scaffolds for Regenerative Medicine. *Annals of Biomedical Engineering* **2015,** *43* (3), 657-680.

Chapter 7: Conclusions and Future Work

7.1 Overall Conclusions

Protein engineering allows us to functionalize proteins of interest for the specific applications required in medicine and industry. Nature has a vast library of proteins of whose functions we can use to our advantage, and protein engineering gives us the tools to optimize them for non-natural applications. Protein engineering allows us to form chimeras, taking parts of different proteins to create new tools. The work described in this dissertation demonstrates how natural proteins can be used as a jumping-off point to design binding proteins through modification of their physical properties at the sequence level to create protein scaffolds for use in non-native conditions.

7.1.1 *H. magnipapillata* NOD-like receptor 42PRs

In this project, we used statistical design to create 42PR proteins based on the presumed recognition domain of *H. magnipapillata* innate immunity receptors. We discovered that these domains were very long and had a high sequence similarity, suggesting the possibility of binding to a polyvalent target. Ultimately this project was met with challenges associated with a multiple sequence alignment that lacked sufficient diversity. We made the choice in building our multiple sequence alignment to include only sequences of known function from *H. magnipapillata*, as we were interested in probing the diversity of innate immunity receptors. Though we had 1285 unique sequences, the structural residues were almost identical from sequence to sequence, with the binding residues responsible for the majority of the sequence diversity. Future work may include a BLAST search to increase the diversity of the structural residues in our multiple sequence alignment library, which could lead to a more stable consensus protein.

7.1.2 Designed Family 1 Cellulose-Binding Module

In this chapter, we used statistical design strategies in order to better understand and improve upon natural cellulose binding modules. Designed CBM1s could improve upon currently proposed synergistic additives to the cellulase cocktail used in enzyme saccharification of lignocellulosic biomass, or be used as a possible domain in designer cellulases. Modifying the surface residues that interact directly with cellulose had a surprisingly large effect on the structure and binding to cellulose. rCBM1-YW had a binding affinity to microcrystalline and amorphous cellulose that matched or exceeded that of naturally occurring CBM1s, while increasing stability.

6.1.3 Adding an Integrin Binding Site to Recombinant Keratins

Biomaterials have benefited from recombinant DNA technology in a number of ways, including the ability to encode features of interest into the primary sequence of the protein. We describe in this chapter an example sequence modification to tune the properties of keratin biomaterials through the addition of an integrin binding motif. The addition of this motif does not affect the proteins' expression or ability to oligomerize as a homo- or heterodimer. We propose that doping recombinant keratin oligomers with modified properties into extracted keratin materials will result in materials with tunable properties for applications in tissue engineering, nerve regeneration, and wound healing.