Quantitative Computational and Biophysical Investigation of Multivalent Proteins

A Dissertation Presented

by

Vadim Patsalo

to

The Graduate School

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

December 2012

Stony Brook University

The Graduate School

Vadim Patsalo

We, the dissertation committee for the above candidate for the
Doctor of Philosophy degree, hereby recommend
acceptance of this dissertation.

David F. Green – Dissertation Advisor
Associate Professor, Department of Applied Mathematics and Statistics

Robert C. Rizzo – Chairperson of Defense
Associate Professor, Department of Applied Mathematics and Statistics

Thomas MacCarthy
Assistant Professor, Department of Applied Mathematics and Statistics

Daniel P. Raleigh
Professor, Department of Chemistry
Stony Brook University

This dissertation is accepted by the Graduate School

Charles Taber
Interim Dean of the Graduate School

Abstract of the Dissertation

Quantitative Computational and Biophysical Investigation
of Multivalent Proteins

by

Vadim Patsalo

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2012

The valency of a biological molecule is the number of interactions that it is able to make with other molecules. Multivalency arises in proteins which oligomerize or contain tandem repeats, and commonly involves the binding of carbohydrates. Biological processes which rely on multivalency include surface interactions (e.g. virus–cell adhesion and cell–cell binding) or de-mixing phenomena. The energetics of multivalent interactions can be significantly enhanced relative to their monovalent counterparts, and numerous multivalent inhibitors of viral infection have been identified. In this dissertation, we present computational and biophysical investigations of several multivalent proteins related to human viral pathogens.

The bivalent lectin Cyanovirin-N inhibits HIV infection by binding the high-mannose glycans on the surface of the viral glycoprotein gp120. We performed Poisson–Boltzmann calculations and identified adjacent serines sequestered in the protein core which form a bridging interaction. We showed that this interaction does not overcome the desolvation penalty for burying the two groups, and went on to design and characterize a series of stabilized protein variants.

The tetravalent lectin MVL also neutralizes HIV, but recognizes a glycan substructure different from Cyanovirin-N. An unresolved question regarding MVL and other HIV-neutralizing agents is whether multivalency is necessary for efficient viral neutralization. We biophysically characterized individual monovalent domains of MVL. The C-terminal domain was folded and populated a monomer/dimer equilibrium at micromolar concentrations. HIV neutralization experiments revealed that the C-terminal domain alone was able to neutralize three of four viral strains with efficacy near that of the wild type protein, suggesting that multivalency is not necessary for nanomolar inhibition by this protein.

The adenoviral protein E4–ORF3 forms a heterogeneous polyvalent nuclear fiber and inactivates several host responses to the infection. Using biophysical techniques we characterized a nonfunctional variant of E4–ORF3, revealing that a homodimer is the building block of the nuclear web. Based on a subsequently-solved X-ray structure, we propose mechanisms of how the mutation abrogates function, and how E4–ORF3 is able to capture a diverse panel of host cellular proteins.

Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to thank my advisors David F. Green and Daniel P. Raleigh for their advice, encouragement, and guidance. Working with David and Dan has been an incredible experience. David has given me the freedom and confidence to pursue my own ideas, has taught me how to solve scientific problems and how to critically think about the scientific literature. Dan let an untrained mathematician into his experimental lab, has been extremely generous with his time and resources, and has provided countless pieces of "fatherly advice" which have shaped my path. From Dan I learned how to design experiments in which one learns something regardless of the outcome, and the "Raleigh way" of running a scientific investigation. Over the years Dave and Dan somehow managed to turn a confused mathematician into an accomplished protein biochemist.

I feel honored to have had the opportunity to work with Patrick Hearing on E4–ORF3, and would like to thank him for listening to my ideas and for his support. Pat introduced me to many concepts in eukaryotic biology, and has shaped my way of thinking.

I would like to thank members of the Green lab, in particular Yukiji K. Fujimoto, with whom I worked closely on computational studies of Cyanovirin-N and MVL. Loretta Au has provided inspiration, especially through her dedication to photo-documenting Green lab events. In the Raleigh lab, I am indebted to Hümeyra Taşkent for helping me get started with experimental studies, and for her friendship. I would like to thank Shifeng Xiao, Wenli Meng, Bowu Luan, Ping Cao for helping me collect a few pieces of experimental data.

Finally, I would like to thank my friends and family for their love and support. I would like to extend my sincerest gratitude to Suzanne Thomas for being by my side throughout this process, and for her love and support. This thesis is dedicated to my mother, who gave me everything I could ever ask for.

Chapter 1

General Introduction

1.1 Multivalent Interactions in Biology

Biology is driven by the non-covalent association of molecules. Among countless examples of noncovalent biomolecular association events are: regulation of protein production by transcription factors, enzymatic binding and turnover of metabolic substrates and cofactors, and the assembly of lipids to form membranes. The valency is a key property of a macromolecule, and is defined as the number of separate ligand–receptor interactions the molecule is capable of simultaneously engaging [52, 119]. Biomolecular interactions (involving proteins, saccharides, nucleic acids, lipids, small molecule metabolites, etc.) or their assemblies (cells, viruses) can be divided into three broad classes based on the valency of the ligand and receptor molecules involved.

1. The first class of interactions includes *uni*valent association (e.g. the binding of bovine pancreatic trypsin inhibitor to trypsin, or the binding of ribosomal proteins to rRNA). Here, a receptor contains a single binding cavity, which can be occupied by a ligand. These interactions have only two states (bound or unbound) and the corresponding equilibrium affinity constant and free energy of binding are derived by measuring the relative populations of these two states (Figure 1-1A).

2. The second class of interactions involves *multi*valent receptors and monovalent ligands. Textbook examples of this class are the binding of molecular oxygen to hemoglobin and of metals to the chelator EDTA. This class of interactions, in certain cases, gives rise to the phenomenon of cooperativity. For instance, the binding of oxygen to hemoglobin is the classical example of positive allosteric cooperativity, in which the binding of the first ligand increases the affinity of subsequent binding events. In general, positive cooperativity can arise from a conformational change or a shift in the receptor equilibrium [85]. Metal chelation is an example of configurational cooperativity, where the initial binding event pre-organizes other receptors (carboxylate groups in the case of EDTA) to engage the ligand. As a class, these interactions are *not* considered multivalent [52].

3. The third class of interactions involves the simultaneous engagement of multivalent receptors with multivalent ligands. Among these are the interactions between surfaces, cells, and macromolecular assemblies. Examples of these include: binding of a virus to receptors on the host cell, adhesion of uropathogenic bacteria to epitheleal cells, or the association of toxins with cell membranes. Multivalent molecules and interactions of this class can exhibit characteristics absent from their monovalent counterparts, and can be significantly enhanced in their strength compared to their monovalent counterparts, reaching levels of effective irreversibility. In general, multivalent interactions can be positively or negatively cooperative, or display no cooperativity at all.

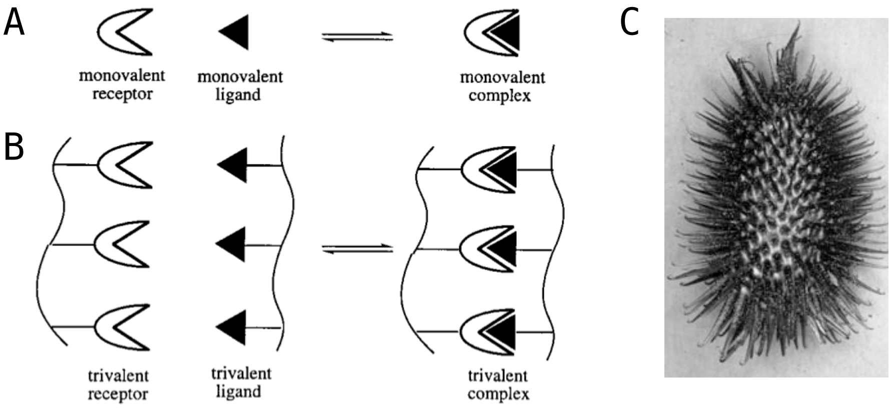


Figure 1-1: **Schematic representation of multivalency.** **A**. Interactions between monovalent receptors and ligands populate a bound/unbound equilibrium, which can be described by an equilibrium constant and an associated binding free energy. **B**. Multivalent interactions present several experimental and theoretical challenges. For large polyvalent systems, determining the receptor fractional occupancy is impossible in many cases. Shown here is a schematic of a trivalent receptor interacting with a trivalent ligand. More complex, heterogeneous binding between these two molecules can also occur, which can lead to large assemblies or aggregates. Panels **A** and **B** adapted from reference [119]. **C**. On the macroscopic level, the *Xanthium* burr can be thought to engage in multivalent interactions. The attachment of the burr to a multivalent receptor (such as an animal's coat or a sweater) involves a large number of simultaneous yet independent noncovalent interactions. The sum of these interactions produces a high burr–sweater affinity which can withstand large shear forces. However, since the interactions are non-covalent, the burr can be detached by sequentially unfastening each hook. Burr image credit: http://www.dpi.nsw.gov.au/agriculture/pests-weeds/weeds/profiles/californian-burr

Multivalency and cooperativity are key features of biological systems, and their presence leads to collective properties which are absent from individual constituents [187]. Even without the affinity gains due to positive allosteric cooperativity, multivalent interactions on the molecular and cellular levels represent a fundamental conceptual departure from their monovalent counterparts. Polyvalent interactions offer numerous advantages, including achieving tight binding using modest surface area, creating contacts between large biological surfaces, and assembly of specific molecular geometries [119]. Multivalent interactions can also give rise to de-mixing phenomena common to eukaryotic organisms; the prototypical example is the formation of “nuclear bodies” by the multivalent protein PML [14]). Recently, a synthetic system has been described which undergoes a de-mixing phase transition similar to those observed in nature [106]. Such phase transitions may be biologically useful in order to spatially organize and regulate biochemical phenomena on the macroscopic scale.

1.1.1 Examples of Carbohydrate-mediated Multivalent Interations

Carbohydrates and carbohydrate–binding proteins are found on the surfaces of cells, viruses, and bacteria; multivalent carbohydrate–protein interactions mediate inter-cellular processes such as adhesion and pathogenesis (Figure 1-2). Understanding the role glycosylation plays in these processes is crucial for the design of successful therapeutic approaches. It has been suggested that significant therapeutic breakthroughs will occur through appreciation and study of multivalent interactions as a whole, and through the incorporation of insights gleaned from these studies to inhibitor design [119].

Viruses begin their life cycles through an attachment to the host cell. This step frequently involves multivalent interactions, and is commonly carbohydrate-mediated. For example, the Human Immunodeficiency Virus (HIV) is decorated with 8–10 heavily glycosilated “spikes” on its surface. These spikes (trimers of glycoproteins gp41 and gp120) target the host receptor CD4 and CCR5/ CXCR4, which are present in multiple copies on the cell surface [45].

Numerous inhibitors of HIV infection have been discovered, and these molecules are frequently multivalent [6]. Among these are naturally-occurring lectins and neutralizing antibodies. Lectins are carbohydrate-binding proteins of non-immune origin which have no enzymatic activity upon their cognate substrates [110]. Lectins are typically multivalent, and are frequently isolated from their original source due to their ability to agglutinate cells or particles which display carbohydrates on their surface (e.g. erythrocytes). Protein–carbohydrate interactions are routinely multivalent because carbohydrate-binding proteins are commonly composed of repeats and/or multimers of modular domains. It is thought that multivalent interactions help many lectins overcome low (mM-range) affinity at individual binding sites [87, 148, 179].

The biological context of a lectin’s function within its native organism is often elusive. This may be due to the unavailability of the genome sequence, or because the lectin is expressed in its native organism under a set of conditions which are difficult to replicate *ex vivo*. Intriguingly, this is often not the main focus of the scientific investigation involving lectins. A number of lectins have found therapeutic or biotechnological use outside their native environment; examples of biotech utility of lectins include: lectin chromatography for glycoprotein purification and lectin microarrays for carbohydrate content profiling of bacteria or viruses [81, 82].

Lectins can interfere with the action of human pathogens, and thus have generated therapeutic interest [6]. However, efforts to engineer improved lectin inhibitors have had limited success, and

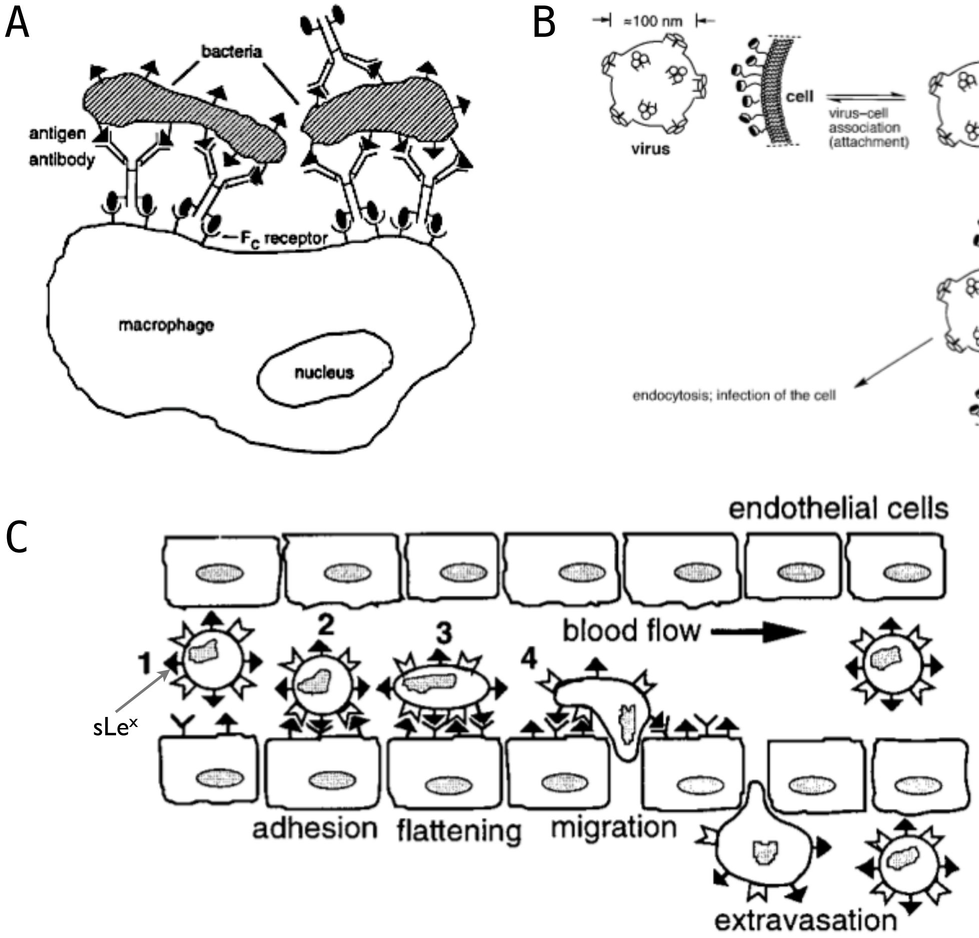


Figure 1-2: **Multivalency on the molecular and cellular levels. A**. The action of the immune system is dependent on multivalency. Divalent IgG antibodies recognize an antigen epitope on the surface of a bacterial cell. The F_C region of antibodies is mannosylated, and interacts multivalently with the F_C receptors present on the surface of macrophages, resulting in two "layers" of multivalent interactions. **B.** Attachment of viruses to cells is mediated by multivalent interactions. Here, the attachment of an influenza virus to the host cell is depicted; the hemagluttinin molecules on the surface of the virus recognize host sialic acid moieties. Viral attachment leads to deposition of the viral material inside the host cell *via* endocytosis. **C.** Injury to endothelial tissue results in enhanced expression of the E- and P-selectins on the cell surface. Neutrophils are attracted to this site, and interact multivalently through the sLe^{x} groups found on the endothelial surfaces. Figure adapted from reference [119].

rely mainly on increasing the number of binding sites through dimerization or tandem duplication. The lack of success is partly due to the fact that the lectin targets are not well characterized. In this work, we present two investigations of multivalent lectins which aim to design improved inhibitors, as well as to understand the role of multivalency in their function. Inhibition of HIV infection by lectins is mediated by the viral surface glycoprotein gp120. Each molecule of gp120 contains approximately two dozen N-linked glycosylation sites, depending on the viral strain, and both high-mannose (for example $Man_9GlcNAc_2$) and complex oligosaccharides decorate its surface [6]. It is thought that the main function of carbohydrates on HIV is to create a “glycan shield”, the primary role of which is to obscure the underlying viral proteins from detection by the host immune system [141, 183].

1.2 Overview of the Present Work

In this work, we describe investigation of three proteins which rely on multivalency for function. The first two studies are of multivalent lectins which neutralize infection by HIV. The final study describes the adenoviral protein E4–ORF3, which inhibits anti-viral host proteins by forming large polyvalent assemblies within the nucleus.

In Chapter 2, we used computational tools to discover a destabilizing interaction within the core of the divalent HIV-inhibitory lectin Cyanovirin-N (CVN). CVN is composed of two homologous domains. The initial design goal of this study was to create a *split*-CVN molecule for “mix-and-match” combinatorial screening of carbohydrate specificity against the many possible carbohydrate epitopes on the viral glycoprotein gp120. However, neither domain containing the wild-type sequence proved possible to produce experimentally. We focused our attention on enhancing the stability of the CVN domains within the context of the wild-type molecule and, using Poisson–Boltzmann electrostatic calculations, discovered a serine–serine interaction buried within the core of CVN domain A. We demonstrated that this serine pair is destabilizing to the protein fold. This is likely because the favorable electrostatic interaction between the two sidechains is inadequate to overcome their cumulative desolvation penalties. We further showed that CVN variants incorporating larger aliphatic substitutions at the serine sites were more stable than the wild-type molecule, and extended these mutants to domain B of the wild-type molecule. The stabilized variants retained both the Cyanovirin fold and its exquisite carbohydrate specificity. In addition, we found that CVN unfolding by the chaotropes urea and guanidinium hydrochloride proceeds on the time scale of days, and is not strictly two-state. Previous studies characterizing the effect of mutations on CVN stability were misled by its slow unfolding, and we provide updated thermodynamic parameters of its denaturation.

Chapter 3 extends the concept of the combinatorial carbohydrate-binding scaffold to the tetravalent HIV-inhibitory lectin MVL. Unlike the topologically-unusual CVN domains, the 54-residue domains of MVL are composed of a ubiquitously-occurring secondary structure motif; as a result, these domains may be more likely to retain their fold when excised from the full-length protein. We found that the homologous amino- and carboxy-terminal domains of MVL ($\mathrm{MVL_N}$ and $\mathrm{MVL_C}$, respectively) differed in their stability. While $\mathrm{MVL_C}$ was folded, $\mathrm{MVL_N}$ was unfolded under all experimental conditions. Despite numerous attempts to obtain folded $\mathrm{MVL_N}$ through rational design and symmetrizing approaches, our efforts were not successful. We confirmed that $\mathrm{MVL_C}$ is dimeric at µM concentrations, and tested the ability of the domain to neutralize viral infection alongside the wild-type protein. The inhibitory potency of the single $\mathrm{MVL_C}$ domain was indistinguishable that of its tetravalent parent, suggesting that multivalency is not strictly necessary for potent antiviral neutralization by MVL.

In Chapter 4, we present the results of a computational investigation into carbohydrate recognition by MVL. Relying on structural modeling and long-timescale molecular dynamics simulations, we investigated the binding of a series of high-mannose and chitin-derived oligosaccharides. We found that $\mathrm{MVL_N}$ is less complementary to high-mannose oligosaccharides than $\mathrm{MVL_C}$. In addition, we observed unprecedented carbohydrate–carbohydrate interactions in the tetravalent wild-type molecule. Finally, our structural modeling, in concert with re-examination of experimental data, uncovered a dual binding mode of the tetrasaccharide $\mathrm{GlcNAc_4}$ to MVL.

In Chapter 5, we applied biophysical techniques to experimentally characterize the adenoviral protein E4–ORF3. Upon viral infection E4–ORF3 forms fibrous aggregates (termed tracks) within the nucleus of the infected cell. A number of host proteins, including the nuclear body-forming PML and its associated transcription factors, subsequently or concurrently relocalize into the tracks, which disables their function or targets them for proteosomal degradation. The exact mechanism of E4–ORF3 function is presently unknown. Nuclear aggregates of E4–ORF3 may be amorphous or amyloid-like in nature, and it is thought that their formation creates a polyvalent scaffold which captures host proteins. The scaffold may function by reducing the entropic costs of substrate binding, or by forming novel emergent protein-binding interfaces. We demonstrate that wild-type E4–ORF3 and non-functional mutant L103A possess identical secondary structure with significant α-helical content, revealing that the mutation does not work by a trivial unfolding mechanism, and suggesting that the nuclear tracks are composed of well-folded proteins. Intriguingly, hydrodynamic characterization revealed that the two variants differ in their association state: while recombinantly-produced wild-type E4–ORF3 is heterogeneously oligomeric (and functional), L103A is predominantly trapped in a homodimeric state (and is non-functional). We hypothesize that the dimer is the building block of the nuclear tracks, and demonstrate that during co-expression, L103A abrogates track formation by wild-type E4–ORF3 in a dominant-negative fashion.

We extend the insights gleaned from our initial E4–ORF3 studies in Appendix A. Relying on a recently-discovered connection between E4–ORF3 and the small ubiquitin-like modifier (SUMO) post-translational modification, we investigated the *in vitro* interaction between E4–ORF3 and human SUMO1, and determined that L103A interacts with SUMO1 weakly, beyond experimental detection. Based on examination of the recently-solved structure of an E4–ORF3 variant we propose a mechanism by which the L103A mutation affects E4–ORF3 oligomerization. We performed bioinformatic analysis of the E4–ORF3 sequence and suggest that the C-terminal domain of E4–ORF3 contains a SUMO-interaction motif (SIM). This previously-undetected putative motif shares remarkable homology with characterized SIMs, and many of the host proteins affected by E4–ORF3 contain the SUMO post-translational modification. Finally, we present a structural model of the E4–ORF3 SIM bound to SUMO1, and discuss this putative complex in the context of known E4–ORF3 orthologs and non-functional mutants.

Chapter 2

Rational and Computational Design of Stabilized Variants of Cyanovirin-N which Retain Affinity and Specificity for Glycan Ligands*

Abstract

Cyanovirin-N (CVN, UniProt ID: P81180) is an 11-kDa pseudo-symmetric cyanobacterial lectin that has been shown to inhibit infection by the Human Immunodeficiency Virus (HIV) by binding to high-mannose oligosaccharides on the surface of the viral envelope glycoprotein gp120. In this work we describe rationally-designed CVN variants that stabilize the protein fold while maintaining high affinity and selectivity for their glycan targets. Poisson–Boltzmann calculations and protein repacking algorithms were used to select stabilizing mutations in the protein core. By substituting the buried polar side chains of Ser11, Ser20, and Thr61 with aliphatic groups, we stabilized CVN by nearly 12 °C against thermal denaturation, and by 1 M of GuaHCl against chemical denaturation, relative to a previously-characterized stabilized mutant. Glycan microarray binding experiments confirmed that the specificity profile of carbohydrate binding is unperturbed by the mutations, and is identical for all variants. In particular, the variants selectively bound glycans containing the Manα(1→2)Man linkage, which is the known minimal binding unit of CVN. We also report the slow denaturation kinetics of CVN and show that they can complicate thermodynamic analysis; in particular, the unfolding of CVN cannot be described as a fixed two-state transition. Accurate thermodynamic parameters are needed to describe the complicated free energy landscape of CVN, and we provide updated values for CVN unfolding.

*Portions of this chapter have been previously published as:

Patsalo V, Raleigh DP, Green DF. Rational and Computational Design of Stabilized Variants of Cyanovirin-N which Retain Affinity and Specificity for Glycan Ligands. *Biochemistry*, 50(49):10698-712. (2011)

2.1 Introduction

The 11 kDa lectin Cyanovirin-N (CVN), originally isolated from the freshwater cyanobacterium *Nostoc ellipsosporum*, has generated interest as a potential anti-viral agent. Therapeutic interest in CVN stems from its ability to potently and irreversibly inactivate both laboratory-adapted and naturally-occurring strains of the Human Immunodeficiency Virus (HIV) [70]. This irreversible inhibition involves tight binding of CVN to the high-mannose oligosaccharides on the viral envelope glycoprotein gp120 [26]. The binding event affects the flexibility of gp120, and is thought to hinder the conformational changes which are required for proper interactions with the cell-membrane receptor CD4 and co-receptors CCR5/CXCR4, which are essential for subsequent gp41-mediated membrane fusion [6, 41].

The presence of glycans on gp120 is crucial to the ability of HIV to evade detection by the immune system. The dense glycan coating of gp120 gives rise to what has been termed a "glycan shield" due to its masking of the underlying immunogenic protein epitopes [183]. In fact, most of the surface of the proteins exposed on the extraviral side of the envelope is covered by carbohydrates. CVN, along with a number of other lectins, represents an example of a novel class of therapeutic carbohydrate-binding agents against enveloped viruses. The antiviral activity of such molecules is dual in mechanism; first, they are able to bind to the glycans of the viral envelope and block virus entry, and second, long-term exposure to such agents leads to a progressive deletion of the glycosylation sites on the viral surface as an adaptive response to antiviral pressure. In the case of HIV, the deletion of the "glycan shield" is thought to reveal previously-obscured epitopes and allow enhanced detection and neutralization of the virus by the immune system [6, 151]. CVN has also been shown to inhibit transmission by other enveloped viruses: in addition to HIV, CVN inhibits Ebola and herpes simplex viral infection by binding to their respective envelope glycoproteins [13, 175], and influenza infection by binding to hemagglutinin [138].

Wild-type CVN has modest stability and a tendency to form domain-swapped dimers [10]. These factors complicate its clinical use and can make biophysical and biochemical studies of the protein difficult. The most promising clinical application of CVN is as a topical microbicide, with hope of reducing HIV transmission in sub-Saharan Africa. The therapeutic must thus be stable under a wide range of conditions and possess a long shelf life under harsh conditions. The interest in developing CVN as a microbicide has led to investigations of large-scale production of the protein using bacterial, yeast, and plant expression systems [35, 126, 159]. Stabilization of CVN has implications for recombinant production of the protein as a therapeutic, and can aid in recombinant protein purification. In addition, availability of stabilized CVN variants facilitates studies of mutations which significantly disturb the stability of the protein fold, such as previously-characterized binding-site knockouts [12, 31, 56, 121, 122] or designed protein oligomers [90].

We set out to design variants of CVN which are more stable to chemical and thermal unfolding, but retain the full biological activity of the wild-type protein. Here, we present CVN variants which are more stable than the wild-type protein, yet retain the native fold and carbohydrate specificity. The designed homologues represent a new background which is amenable to binding-site redesign or engineering, due to their increased tolerance of higher temperatures and denaturant concentrations compared to wild-type CVN.

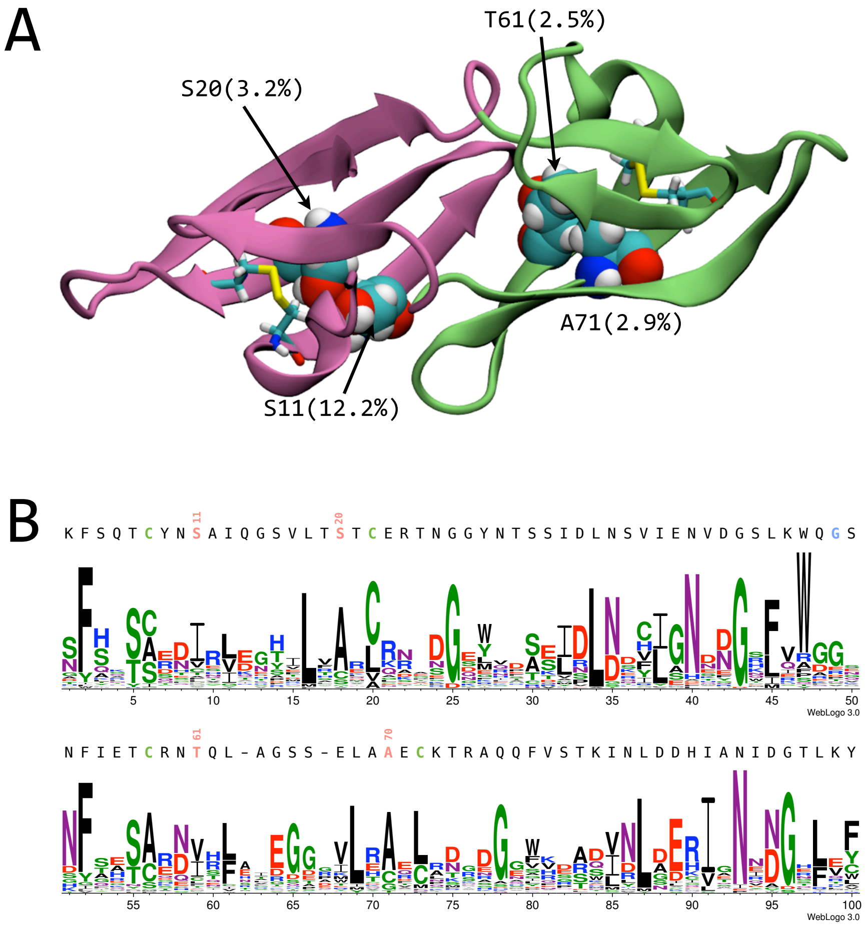

A
T61(2.5%)
S20(3.2%)
A71(2.9%)
S11(12.2%)
B
11
20
K F S Q T C Y N S A I Q G S V L T S T C E R T N G G Y N T S S I D L N S V I E N V D G S L K W Q G S
5 10 15 20 25 30 35 40 45 50
WebLogo 3.0
61
70
N F I E T C R N T Q L - A G S S - E L A A E C K T R A Q Q F V S T K I N L D D H I A N I D G T L K Y
55 60 65 70 75 80 85 90 95 100
WebLogo 3.0

Figure 2-1: **A**. Ribbon representation of Cyanovirin-N based on PDB `1IIY` [15]. Domain A is colored purple, and Domain B green. Disulfide bonds are shown in licorice, and Ser11, Ser20, Thr61, Ala71 are shown in space-filling representation. Fractional side-chain solvent accessibility of these residues is indicated in parentheses. Solvent accessibility was computed with Naccess [83] **B**. Sequence logo representation of CVN family rendered with WebLogo 3.0 [37]; the height of the letter stack at each position is proportional to sequence conservation, while the height of the letters within a stack is scaled to relative amino-acid frequency. The numbers below the logo are position indices in the sequence alignment. The first and second repeats are aligned to emphasize the tandemly-repeated nature of CVN. The sequence of P51G-CVN is shown above.

Protein stabilization has previously been achieved using a variety of approaches: improving core packing [38], removal of buried polar side chains or unsatisfied salt bridges [20,75], homology-based strategies [125], mutation of charged surface residues [171], introduction of new disulfide bonds [147], turn redesign [132,144,199], modulation of unfolded state entropy [3], and rational considerations of unfolded state interactions [34]. Often, the substitutions increase stability via a mixture of effects, such as optimized core packing and increased burial of hydrophobic surface area, which may be difficult to deconvolve [118]. In this work, we employed a rational design strategy which removes buried polar groups and improves the packing within the protein core to yield increased stability.

We found that kinetic denaturation of CVN by guanidine hydrochloride (GuaHCl) is very slow, and that CVN folding is not two-state. Previous studies of CVN folding thermodynamics were misled by the slow unfolding, and we provide updated thermodynamic parameters for CVN.

2.2 Materials & Methods

Continuum Electrostatic Calculations. Electrostatic contribution of CVN side chains to protein stability were obtained using standard methods [67] by solving the linearized Poisson–Boltzmann equation [61, 80] using a multigrid finite-difference solver (M.D. Altman and B. Tidor, unpublished) distributed with the Integrated Continuum Electrostatics (ICE) software package (DFG, E. Kangas, Z.S. Hendsch, and B. Tidor, available for licensing through the MIT Technology Licensing Office). Dielectric constants of 2 and 80 were used for the solute and solvent, respectively. The dielectric boundary was defined by the molecular surface using a 1.4 Å radius probe, with radii optimized for this purpose [135]. The ionic strength was set to 145 mM, with a 2.0 Å ion-exclusion layer. A 129-unit grid was used with overfocusing boundary conditions (the longest dimension of the molecule occupying 23%, then 92%, and finally 184% of one edge of the grid).

The electrostatic contribution of a side chain at position i to the unfolded state was modeled by its interactions with the $(i-1)_{\mathrm{carbonyl}}$-$(i)$–$(i+1)_{\mathrm{amino}}$ "tripeptide" in the absence of any other protein groups. The conformation of the "tripeptide" was unchanged from that of the folded protein. Electrostatic calculations were performed on 201 snapshots extracted from a 200 ns explicit-solvent molecular dynamics simulation of CVN [57].

Protein Design Calculations. We used a hierarchical design procedure [66] based on the Dead-End Elimination and A* algorithms [40,64,99,102,113,117] to find low-energy sequences compatible with the CVN fold. The protein backbone was kept fixed, and the "penultimate" rotamer library of Richardson and colleagues [114] was used for side-chain rotamers. Energies were evaluated with the CHARMM potential [27] with the PARAM22 force field [116] using the analytical continuum electrostatic model [152].

All CHARMM energy terms (bond, angle, dihedral, improper, Lennard-Jones, and electrostatic) were used in the search. Changes to protein stability upon mutation or side-chain rearrangement were approximated from the energetic difference between the folded and the unfolded states based on isolated model side chains. The isolated side chains were acetylated at the N-terminus and N-methylamidated at the C-terminus. Rotamers which clashed with the backbone or with neighboring side chains were eliminated. Sequences greater than 15 kcal mol^{-1} above the global minimum energy configuration were discarded. Ten top-scoring configurations for the remaining sequences were reranked using Analytical Continuum Electrostatics [153], as implemented in CHARMM. Software written in collaboration with Tidor and colleagues [66,109] (available for licensing through the MIT Technology Licensing Office) was used for the search.

Cloning, Protein Expression & Purification. Mutants were constructed starting with the synthetic gene coding for Cyanovirin-N (DNA 2.0) cloned into the pET-26b(+) vector (Novagen) using the NdeI and XhoI restriction sites. Round-the-horn site-directed mutagenesis (S. Moore, unpublished) was performed using the Phusion High-Fidelity PCR Kit. Briefly, non-overlapping primers, with one encoding the desired mutation, were phosphorylated at the 5' end using T4 Polynucleotide Kinase. The phosphorylated primer mix was then added to the PCR reaction, and extension followed for 30–35 cycles. The purified PCR product was then ligated at 16 °C overnight with T4 DNA Ligase, and transformed into XL1-Blue Competent Cells (Novagen). A typical transformation yielded 50–200 colonies. Restriction enzymes, Phusion polymerase, polynucleotide kinase, and ligase were purchased from New England Biolabs. The identity of the mutants was confirmed by DNA sequencing.

Plasmids bearing the appropriate mutations were transformed into BL21(DE3) *E. coli* strain (Novagen). Cells were grown at 37 °C until OD_{600} reached 0.8, and protein expression was induced by addition of 1 mM isopropyl-D-thiogalactoside for 4–5 h. The cells were pelleted by centrifugation at 7500 g, and frozen at −80 °C until purification.

With the exception of ΔM, the proteins were purified under denaturing conditions. The cell pellet was resuspended in Buffer A (6 M GuaHCl, 20 mM imidazole, 20 mM Tris-HCl, pH 8.0) using 10 mL per gram of cell paste. The cells were disrupted by four passes through a French Press high-pressure homogenizer. The insoluble fraction was immediately pelleted by ultracentrifugation at 100 kg. The supernatant was loaded onto a 5ml His-Trap FF column (GE Healthcare), connected to an AKTA Explorer 10 FPLC (GE Healthcare), and eluted over 10 column volumes using a gradient of Buffer B (6 M GuaHCl, 300 mM imidazole, 20 mM Tris-HCl, pH 8.0). Dithiothreatol was added to a final concentration of 5 mM.

The proteins were refolded by overnight dialysis against Buffer C (10 mM Tris-HCl, pH 8.0) at room temperature, changing the buffer once. The precipitate was removed by ultracentrifugation, and the soluble fraction was incubated at 37 °C for 24–48 h to increase interconversion of domain-swapped dimer to monomeric protein. The proteins were concentrated by centrifugation, and loaded onto a Superdex 75 26/60 gel filtration column (GE Healthcare) pre-equilibrated with Buffer D (20 mM sodium phosphate, pH 6.0). The monomeric proteins were stored at 4 °C until characterization.

The identity and purity of the recombinant proteins was confirmed by matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry; all of the overexpressed proteins were found to have an N-terminal methionine residue. Protein concentrations were determined by measuring A_{280} in 6 M GuaHCl using a calculated extinction coefficient of 10 220 $M^{-1}\,cm^{-1}$.

Oligomerization State Determination. Analytical gel filtration was performed on a Superdex 75 (10/300 GL) column (GE Healthcare). 100 µL samples were injected onto the column pre-equilibrated in 20 mM sodium phosphate, 200 mM sodium chloride (pH 7.5) and were eluted in the same buffer at 0.5 mL min^{-1}.

Equilibrium Denaturation Studies. Samples for equilibrium denaturation studies were prepared as follows. Typically, thirty-two 2 mL samples containing 10 µM protein in buffer D with the appropriate concentration of GuaHCl were incubated at room temperature for 72 h. The concentration of GuaHCl in each sample was determined by refractometry. Samples were analyzed by fluorescence and circular dichroism spectroscopies, described below.

Spectroscopic Analysis. Intrinsic tryptophan fluorescence spectra were collected at 25 °C on a PTI spectrofluorometer (Birmingham, NJ) equipped with a Peltier temperature controller. Protein concentrations were 10 µM in Buffer D. The excitation wavelength was set at 290 nm and fluorescence emission was collected from 300–400 nm. To facilitate comparison with earlier work, the intrinsic tryptophan fluorescence ratio I_{330}/I_{360} was also recorded for each sample for 60 s, and averaged to yield the final value.

Circular dichroism measurements were carried out on a Chiroscan spectrometer (Applied Photophysics) equipped with a Peltier temperature controller. Far-UV spectra were recorded from 190 nm–260 nm in 0.5 nm steps using a path length of 1 mm. Spectral measurements were performed with a bandwidth of 2 nm, with a data acquisition length of 1 s. For thermal denaturation studies, protein concentrations were 15 µM in Buffer D. The sample was heated at 2 °C min^{-1} in steps of 1 °C from 5 °C to 90 °C. Thermal unfolding was measured by following the transition at 200 nm, with a data acquisition time of 5 s at each temperature. θ_{200} was chosen for consistency with previous work and because it shows near-maximal signal difference upon heat denaturation.

Thermal denaturation curves were fit to a two-state model [68, 69] using a custom nonlinear least-squares regression routine, implemented in the statistical environment R [146]. Both the pre- and post-transition regions were allowed a linear baseline. Equilibrium denaturation far-UV spectra in the presence of GuaHCl were collected from 215–245 nm in 0.5 nm increments using 10 µM protein and a path length of 5 mm. The data analysis is discussed below.

NMR spectra were collected at 293 K on a Varian Inova 600 MHz spectrometer. All samples contained 150 µM protein in 20 mM sodium phosphate pH 6.0, 10% D_2O. DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid) was included as an internal reference. Spectra were processed with NMRPipe [39], and analyzed with SPARKY (T.D. Goddard and D.G. Kneller, SPARKY 3, University of California, San Francisco)

Glycan Array Screening. Fluorescein-labeled samples of CVN variants were prepared by incubating 0.5 mg purified protein in 50 mM borate, pH 8.5 with a 15-fold molar excess of NHS-Fluorescein (0.3 mg) at room temperature for 1 h. Unreacted dye was removed by desalting on a HiTrap 5 mL column (GE Healthcare), and samples were further dialyzed against PBS overnight in the dark at 4 °C. Protein concentrations were determined by measuring A_{280} and A_{494}, correcting for fluorophore absorbance. Samples were submitted to the Consortium for Functional Glycomics, where 70 µL of 200 µg mL^{-1} solution were spotted onto Mammalian Printed Array Version 4.1 containing 465 carbohydrates.

Data Analysis. GuaHCl-induced denaturation data were fit as follows. The decomposition of spectroscopic data into basis states and their fractional populations can be represented as a matrix multiplication. CD spectroscopy obeys Beer's law (*i.e.* the observed data is a linear combination of the basis spectra weighted by their fractional populations). Spectra of m wavelengths recorded over n experimental conditions (denaturant concentrations) populate an $m \times n$ matrix A [78]. Decomposition of these data into k thermodynamic states (in our case, $k = 3$) is equivalent to the following matrix product:

$$A_{m\times n} = S_{m\times k}F^T_{n\times k}$$

The columns of matrix S contain the spectra of the thermodynamic states, while the columns of F^T are the fractional populations of the thermodynamic states at each experimental condition. The data are obscured by an $E_{n\times m}$ matrix of experimental noise. Assuming the linear extrapolation (LEM) model, in which the free energy of denaturation is linear with denaturant concentration [131], the relative free energies of the thermodynamic states at any denaturant concentration are given by four thermodynamic parameters ($\Delta G^\circ_{\mathrm{N}\to\mathrm{I}}$, $m_{\mathrm{N}\to\mathrm{I}}$, $\Delta G^\circ_{\mathrm{N}\to\mathrm{D}}$, $m_{\mathrm{N}\to\mathrm{D}}$), which are defined in Table 2.4. If the relative free energies of the three states at a particular denaturant concentration are known, their fractional populations are given by the partition function, and the matrix F is fully specified. The matrix S is then obtained by taking the pseudoinverse:

$$S = A(F^T)^+$$

The goal is to minimize the error between the experimental data A and the model SF^T. The objective function in our model is as follows:

$$\left|\left|A - SF^T\right|\right|_2 + \alpha||S||_2$$

The operation $||\cdot||_2$ is the matrix 2-norm. Our objective function is a combination of an error term and a weighted ($\alpha = 0.01$) regularization term. The inclusion of a regularization term was necessary in order to obtain physically meaningful solutions. Without regularization the optimization returns spectra which can be nearly infinite in magnitude, while low in fractional population. The regularization weight was set at 0.01, as this value allowed the accurate recovery of thermodynamic parameters in a synthetic data set. The objective function was minimized using the Nelder–Mead simplex algorithm, as implemented in the `optim()` function of the statistical environment R [146].

2.3 Results

Cyanovirin-N is composed of two tandem repeats of 50 and 51 amino acids which share 32% sequence identity (Figure 2-1A) [15]. These repeats give rise to two pseudo-symmetric domains, termed Domain A (residues 1–39 and 90–101) and Domain B (residues 40–89). Each domain contains a carbohydrate-binding site [19] and is stabilized by an internal disulfide bond. The two domains make extensive contacts and share a hydrophobic core. Despite the structural pseudo-symmetry of the two domains, Domain A is not contiguous in sequence, and contains both amino- and carboxy-termini, while the structurally-equivalent residues within Domain B are connected by a short linker (Lys48–Trp49–Gln50–Pro51–Ser52–Asn53).

We designed mutants using the Pro51Gly (P51G) mutation as background. This mutation, located in the linker region of Domain B, has previously been shown both to stabilize monomeric CVN and to increase the yield of monomeric protein relative to the domain-swapped dimer, a metastable state which is formed when the protein is refolded at high concentrations [10]. Wildtype CVN lacking this mutation forms a domain-swapped dimer under crystallographic conditions. The crystal structure of domain-swapped CVN shows significant rearrangement in the linker region, and exhibits changes in the ϕ and ψ torsion angles, which are most profound for Ser52 and Asn53 [197]. The P51G substitution is thought to stabilize CVN by alleviating backbone strain induced by Pro51, which imparts a positive ϕ angle on Ser52 and places it into a disfavored portion of the Ramachandran plot. In molecular dynamics simulations of wild-type CVN, Fujimoto *et al.* observed significant rearrangements in this linker region, including formation of a *cis*-peptide bond between Pro51 and Ser52 [57].

2.3.1 Ser11, Ser20, and Thr61: Buried Polar Targets for Rational Design

Initial insight into stabilizing CVN came from the observation that Domain A side chains of Ser11 and Ser20 are buried within the hydrophobic core of the protein, with solvent accessibility of 3.3% and 12.2%, respectively, relative to the Ser side chain in an Ala–Ser–Ala tripeptide (Figure 2-1B). While the average NMR structure (PDB ID 1IIY) does not show a hydrogen bond between these groups, the interaction is nearly always formed in molecular dynamics simulations of CVN.

In proteins, polar groups are more frequently found on the surface than in the core and often pay significant desolvation penalties upon burial [25]. Thr61 and Ala71, which are the Domain B symmetric counterparts of Ser11 and Ser20, pack against one another, are within van der Waals contact of Phe54, and contribute to the hydrophobic core of the protein. These packing interactions are absent in Domain A, despite the proximity of Phe4 (the symmetric equivalent of Phe54) to Ser11 and Ser20 [17].

Ser11 and Ser20 are among the least solvent-exposed side chains in the protein. The side chains of these two residues appear to make a direct hydrogen bond in the solution structure of CVN, and this favorable interaction may compensate for the desolvation penalty they pay for burial within a hydrophobic environment [75, 76]. Thr61, like Ser11, is also excluded from solvent, and only exposes 2.5% of the side-chain surface area, when compared to a reference tripeptide. Unlike its symmetry-related counterpart, Thr61 does not form a buried hydrogen bond. We hypothesized that replacement of Ser11, Ser20 or Thr61 with an appropriate hydrophobic isostere would stabilize the protein, as replacement of Ser with Ala, and of Thr with Val, was previously shown to be favorable in buried positions within T4 lysozyme [20].

2.3.2 Poisson–Boltzmann Calculations Reveal a Destabilizing Polar Bridge

We performed Poisson–Boltzmann (PB) continuum electrostatic calculations in order to estimate the contribution of these groups to protein stability. The calculations afford the desolvation penalty and the strength of the pairwise electrostatic interactions between the side chains of our target positions (Ser11, Ser20, and Thr61) and other groups in the protein. Ala71, the symmetry-related partner of Ser20, was also included as a control. The calculations evaluate the effect of substituting a side chain with a hydrophobic isostere having the same size and shape, but devoid of charge [75,76].

Domain A side chains of Ser11 and Ser20 pay a +2.92 and +2.96 kcal mol^{-1} desolvation penalty ($\Delta\Delta G^\circ{}_{\text{solv}}$), respectively, upon burial within the core of CVN (Table 2.1). For each side chain, the desolvation penalty is opposed by the favorable electrostatic interactions with the other groups in the protein. These interaction free energies ($\Delta\Delta G^\circ{}_{\text{inter}}$) sum to −4.25 and −3.39 kcal mol^{-1} for Ser11 and Ser20, respectively. The major portion of these interactions comes from the direct interaction between the two groups of −3.08 kcal mol^{-1}.

Molecular group	$\Delta\Delta G^\circ{}_{\text{solv}}$	$\Delta\Delta G^\circ{}_{\text{inter}}$	$\Delta\Delta G^\circ{}_{\text{mut}}$	$\Delta\Delta G^\circ{}_{\text{inter.partner}}$
Ser11	$+2.92 \pm 0.02$	-4.25 ± 0.09	-1.33 ± 0.08	-3.08 ± 0.06
Ser20	$+2.96 \pm 0.02$	-3.39 ± 0.08	-0.43 ± 0.08	-3.08 ± 0.06
Ser11.Ser20	$+5.88 \pm 0.03$	-4.56 ± 0.09	$+1.32 \pm 0.09$	
Thr61	$+2.30 \pm 0.02$	-2.51 ± 0.08	-0.21 ± 0.08	$+0.03 \pm 0.01$
Ala71	$+0.03 \pm 0.01$	$+0.20 \pm 0.01$	$+0.23 \pm 0.01$	$+0.03 \pm 0.01$
Thr61.Ala71	$+2.33 \pm 0.02$	-2.34 ± 0.08	-0.01 ± 0.08	

Table 2.1: **Calculated electrostatic contribution of CVN amino-acid side chains to protein stability.** Values were calculated for 201 molecular dynamics snapshots and averaged. The standard errors of the mean are provided as a measure of uncertainty. All free energy values are in kcal mol^{-1}.

In order to evaluate whether a polar group or its hydrophobic isostere is more favorable to the stability of a protein fold, we calculate the mutation free energy $\Delta\Delta G^\circ{}_{\text{mut}}$, which is tabulated as the sum of $\Delta\Delta G^\circ{}_{\text{solv}}$ and $\Delta\Delta G^\circ{}_{\text{inter}}$. In our model, $\Delta\Delta G^\circ{}_{\text{mut}}$ corresponds to "turning on" the partial charges on the molecular group of interest. A negative value of $\Delta\Delta G^\circ{}_{\text{mut}}$ thus indicates that a group contributes more favorably in the charged state than in the hydrophobic state. For Ser11 and Ser20, an isosteric substitution is unfavorable by +1.33 and +0.43 kcal mol^{-1}, respectively (Table 2.1). A single isosteric substitution is unfavorable at either position because the remaining charged

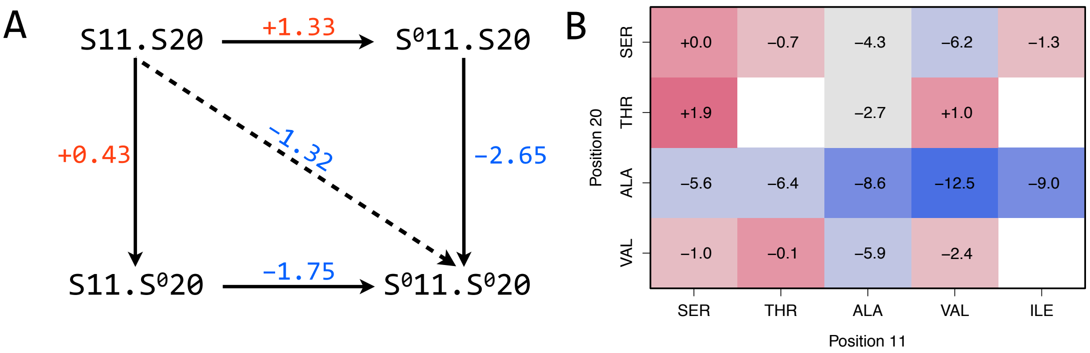


Figure 2-2: **A**. Calculated effect of replacing the side chains of Ser11 and Ser20 with hydrophobic isosteres. All values shown are $\Delta\Delta G°_{\mathrm{mut}}$, in kcal mol^{-1}, with details provided in Table 2.1. **B**. Heatmap summary of protein design calculations on Domain A of CVN. Amino acid substitutions as positions 11 and 20, as well as $\Delta\Delta G°$ values for each sequence are shown. The $\Delta\Delta G°$ values are referenced to wild type (Ser11.Ser20). Sequences corresponding to the entries colored in white were determined as "dead ends" in the design procedure.

group has lost a significant electrostatic interaction partner, yet still pays a desolvation penalty for burial. However, the *simultaneous* replacement of Ser11 and Ser20 with hydrophobic isosteres is predicted as favorable with an estimated effect of −1.32 kcal mol^{-1}. Figure 2-2A shows the complete *in silico* thermodynamic cycle which follows the replacement of either or both serine side chains with hydrophobic isosteres (denoted S^0).

We compared the magnitude of the electrostatic interactions to that of van der Waals interactions experienced by Ser11 and Ser20 using the coordinates taken from a molecular dynamics simulation of CVN. While the electrostatic interactions of the two side chains with the rest of the protein are favorable, the hydrogen bond between these two groups results in overlap of their van der Waals radii, and an unfavorable direct interaction of +0.85 kcal mol^{-1}. These results suggest that the direct interaction between the side chains of Ser11 and Ser20 is mainly electrostatic in nature.

In Domain B, we found that Thr61 pays a desolvation penalty of +2.30 kcal mol^{-1}. This penalty is computed to be nearly perfectly offset by favorable electrostatic interactions with the rest of the protein of −2.51 kcal mol^{-1}, and the overall effect of replacing Thr61 with a hydrophobic isostere is only slightly unfavorable by +0.21 kcal mol^{-1} (Table 2.1). As expected, the nonpolar Ala71 does not participate in significant electrostatic interactions with Thr61.

The incorporation of an exact hydrophobic isostere for Ser into CVN is not possible. Among naturally-occurring amino acids, Ala is the most conservative nonpolar substitution for Ser and is the closest to an isostere. We thus designed a mutant termed AATA. In our nomenclature, a protein variant is identified by a four-letter name which designates its amino-acid identity at positions 11, 20, 61, and 71. For example, P51G variant with no other changes would be denoted SSTA,

Variant	51	11	20	61	71
wild-type	Pro	Ser	Ser	Thr	Ala
P51G (SSTA)	Gly	Ser	Ser	Thr	Ala
AATA	Gly	Ala	Ala	Thr	Ala
VATA	Gly	Val	Ala	Thr	Ala
IATA	Gly	Ile	Ala	Thr	Ala
VAVA	Gly	Val	Ala	Val	Ala
IAIA	Gly	Ile	Ala	Ile	Ala

Table 2.2: **CVN variants discussed in this work.** Numbered columns show the amino acid identity at that position. All CVN variants characterized in this work are made in the background of the P51G stabilizing mutation [10].

and AATA is P51G with Ser11Ala and Ser20Ala mutations (Table 2.2). We next employed protein design calculations to determine if other combinations of naturally-occurring amino acids could lead to increased stabilization of CVN.

2.3.3 Design of a Stabilizing Network of Mutations

The PB analysis provides information about isosteric substitutions, but does not consider changes in packing which may result from non-isosteric mutations. Thus we went on to examine potential effects of varying the size and shape of the side chain at positions 11 and 20. In order to determine if the structure of CVN can accommodate other amino-acid side chains at positions 11 and 20, we used a computational protein design algorithm where the design positions were allowed to simultaneously vary in sequence to {Ala, Ser, Thr, Val, Phe, Leu, Ile}. These substitutions were selected due their nonpolar nature or their occurrence within wild-type CVN at the chosen positions. Neighboring residues Phe4, Leu18, Ile34, Leu36, Leu87, Ile91, and Leu98 were allowed to, if necessary, adopt a different rotamer in order to accommodate larger side chains at the design positions. The protein backbone, as well as the remaining side chains, was kept fixed. The design algorithm computed the stability of a conformational arrangement of amino-acid side chains relative to an unfolded state approximation in which the free energy of the unfolded polypeptide is taken as a sum of energies of the individual amino-acid residues in its primary structure [66, 109].

The computational design calculations suggested that a number of mutants would be more stable than the wild-type protein (Figure 2-2B). In particular, S11V.S20A, S11I.S20A, and S11A.S20A were calculated to be more stable than wildtype by -12.5, -9.0, and -8.6 kcal mol^{-1}, respectively. The results suggest a clear preference of alanine to serine at position 20, with X11.S20 always worse energetically than X11.A20 for every substitution X considered. The insights from the protein design calculations led to mutants VATA and IATA (Table 2.2). It is worth noting that in interpreting these results the energetic trends are more meaningful than magnitudes. The computational procedure uses a discrete representation of configurational space and employs an approximate energy function which ignores important contributions to protein stability (such as side chain configurational entropy). Thus, we do not expect the magnitudes of stabilization from computational design to be quantitatively predictive, but the calculations do reveal potentially stabilizing replacements and complement the PB analysis.

2.3.4 Lessons from CVN Homologues

When CVN was originally discovered, it showed no sequence similarity to any other protein. The structure of CVN revealed a novel fold, which possessed only distant (domain-level) topological similarity to known protein folds. More recently, CVN homologues have been discovered in other prokaryotes [91] and in eukaryotes [142]. The Pfam database (release 24.0) lists 116 putative CVN domain sequences across 22 species [54]. These genomic data suggest that the CVN domain is a module which often exists within larger multidomain proteins, the function of which is presently unknown [142]. There is evidence, however, that these domains adopt structures similar to wild-type CVN and are functional lectins. The solution structures of CVN homologues from *Tuber borchii*, *Ceratopteris richardii*, *Neurospora crassa*, *Microcystis aeruginosa PCC7806*, and *Magnaporthe oryzae* have recently been characterized; all adopt the same fold [94, 95, 161].

A SeqLogo [155] representation of a multiple sequence alignment of a subset of CVN homologues is shown in Figure 2-1B. This sequence alignment revealed that across the CVN family, the position corresponding to Ser11 is frequently Val or Ile, while the equivalent of Ser20 is most commonly Ala. Thr61, the Domain B symmetric equivalent of Ser11, is often substituted with Val or Ile. Unfortunately, the size of the CVN family is not yet large enough to determine whether substitutions at these or other positions within the CVN domain are independent or correlated in their conservation [112, 133].

The sequence conservation data are consistent with both computational and intuition-based insights, and reveal a strong preference for nonpolar side chains at the core positions 11, 20, 61, and 71. We thus designed additional symmetrizing mutants, VAVA and IAIA, in order to incorporate the hydrophobic side chains of Val and Ile at position 61 (Table 2.2) in the background of putative stabilizing mutations in Domain A. The distance between the design positions in the two domains of CVN led us to hypothesize that the effect of substitutions at position 61 in Domain B would be additive with the effect of substituting positions 11 and 20 in Domain A.

2.3.5 Expression Construct: Consistency with Previous Work

The biophysical, structural, and inhibitory properties of CVN have been characterized by a number of different laboratories, with some variation in the exact amino-acid sequence of the protein being studied, in particular at the N- and C-terminal regions. The first source of variation is the identity of the N-terminal amino acid. When the protein is expressed cytoplasmically in the BL21 *E. coli* strain it accumulates in inclusion bodies. Consequently, the N-terminal Met residue cannot be processed by the *E. coli* methionine aminopeptidase (confirmed by mass spectrometry) and is retained on the polypeptide chain. In order to facilitate disulfide bond formation, CVN has also been successfully expressed as a folded protein in the periplasmic space, and the removal of the localization tag leaves Leu as the N-terminal residue [127, 128]. In addition, other biophysical studies used CVN variants with an additional N-terminal Gly–Ser–His–Met–Gly sequence which remained after thrombin cleavage. CVN containing these additional five animo acids at the N-terminus exhibited anti-HIV activity which is indistinguishable from wild-type protein [10, 11, 163], yet may have thermodynamic properties different from wild type.

Due to the proximity of the Glu101 side chain to the N-terminus in the solution NMR structure of CVN, it is likely that a salt bridge is formed between these two groups at the experimental pH of 6.0. We hypothesized that the addition of a methionine or a longer linker at the N-terminus could thus have an effect on protein stability by altering the salt bridge geometry, or by interfering with its formation. To evaluate this, we expressed the P51G variant in the *E. coli* periplasm using the PelB localization tag. This variant (termed ΔM) is identical in sequence to P51G, except it is one residue shorter due to the removal of the leader tag.

A second source of variation in the sequence of CVN is the presence of a C-terminal His-tag (complete sequence Leu–Glu–His_6). The thermodynamic parameters of non-His-tagged and His-tagged wild-type CVN are identical [126], and a number of additional studies have used this construct [71]. For ease of purification, our variants contain this C-terminal His-tag.

Variant	T_m (°C)	C_m^{Fl}(M)	m^{Fl}	C_m^{CD}(M)	m^{CD}
P51G	64.3 ± 0.1	2.3 ± 0.02	2.6 ± 0.05	2.2 ± 0.02	2.5 ± 0.09
AATA	69.6 ± 0.3	2.6 ± 0.01	2.5 ± 0.05	2.5 ± 0.02	2.4 ± 0.14
VATA	69.8 ± 0.2	2.8 ± 0.01	2.6 ± 0.06	2.8 ± 0.02	2.6 ± 0.21
IATA	71.1 ± 0.1	3.1 ± 0.01	2.3 ± 0.05	3.1 ± 0.02	2.0 ± 0.11
VAVA	74.5 ± 0.2	3.0 ± 0.01	2.1 ± 0.06	2.9 ± 0.03	1.7 ± 0.12
IAIA	75.8 ± 0.3	3.5 ± 0.01	2.1 ± 0.06	3.3 ± 0.02	1.7 ± 0.09
ΔM	66.6 ± 0.2	2.6 ± 0.02	1.9 ± 0.11	2.5 ± 0.02	2.0 ± 0.10

Table 2.3: **Apparent equilibrium denaturation parameters.** The parameters are derived from direct fitting of circular dichroism and fluorescence data. The data were fit to a simple two-state model, as discussed in the text. The standard errors of the fit are provided as measures of uncertainty.

2.3.6 Designed Variants Adopt the Wild-type Fold and are More Stable

Designed CVN homologues were expressed recombinantly in *E. coli* and (with the exception of ΔM) purified from inclusion bodies under denaturing conditions. The proteins possess identical secondary and tertiary structure, as judged by far- and near-ultraviolet CD spectroscopy (Figure 2-3). The far-UV CD spectra of the variants were similar to previously-published spectra for wild-type CVN [11]. In addition, the intrinsic tryptophan fluorescence emission for each variant was blue-shifted 20 nm compared to the fluorescence of GuaHCl-denatured protein, indicating the burial of the unique Trp in CVN (Trp49). All variants were judged to be monomeric by analytical gel filtration chromatography at a range of concentrations up to 141 µM. In order to further confirm conservation of the CVN fold upon mutation, we collected HSQC spectra of P51G and of IAIA. The spectra (Figure 2-4) show comparable resonance dispersion in ^{1}H and ^{15}N dimensions for both proteins, and are consistent with a well-folded structure.

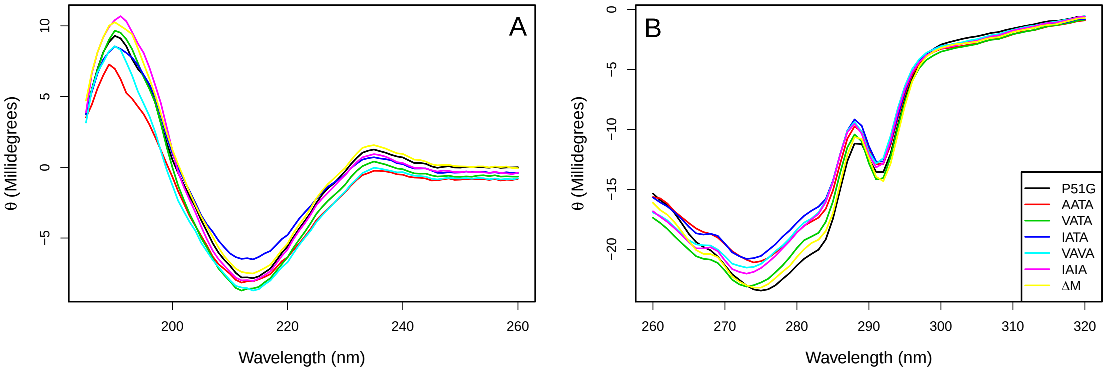


Figure 2-3: Far-(**A**) and near-UV (**B**) circular dichroism spectra of CVN mutants confirm mutants possess identical secondary and tertiary structure. Far-UV spectra were recorded at 10 μM protein concentration using a 1 mm path length cuvette. Near-UV spectra were recorded at 100 μM in a 10 mm path length cuvette. Slight discrepancies of signal intensity are attributed to errors in determining protein concentration.

Figure 2-5 shows the thermal denaturation of CVN homologues monitored by CD spectroscopy, and the extracted thermodynamic parameters are given in Table 2.3. The parameters determined from thermal denaturation represent apparent values, since thermal denaturation of CVN is not fully reversible, as judged by recovery of CD signal upon heating and re-cooling. For some variants, such as P51G, the signal recovery was close to 90%. However, the denaturation of the more hydrophobic variants was less reversible, and visible aggregation within the cuvette was observed.

For P51G, the midpoint of thermal denaturation was 64.3 °C, consistent with previous studies of His-tagged CVN variants [126]. Each of the designed variants were more thermostable than P51G, with thermal stabilization ranging from 5.2 °C for AATA to 11.5 °C for IAIA. In general, greater stabilization was achieved by larger amino-acid substitutions at the design positions. In particular, AATA was less thermostable than VATA, which in turn was less thermostable than IATA. The ΔM variant was more thermostable than P51G by 2.3 °C.

In order to further characterize the stability of the designed variants, we undertook unfolding studies using the chaotrope guanidine hydrochloride (GuaHCl). To facilitate comparison with earlier work, we followed denaturation using the ratio of the intensity of fluorescence emission at 330 nm and at 360 nm, I_{330}/I_{360}. We observed that P51G and other CVN variants look longer than 48 h to fully unfold when exposed to moderate concentrations of GuaHCl at room temperature. For example, incubation of P51G overnight with increasing concentrations of GuaHCl yielded a biphasic denaturation profile. In contrast, complete equilibration could be achieved over the course of 72 h, and at equilibrium, a single sigmoidal transition was observed (Figure 2-6A). This is an

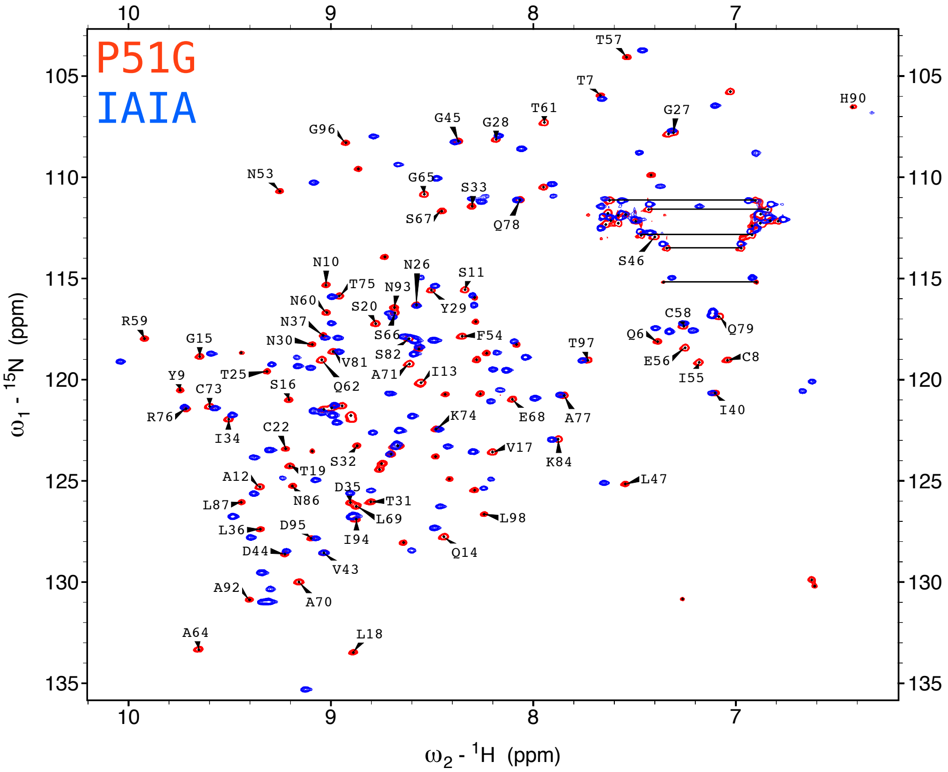


Figure 2-4: Overlay of HSQC correlation spectra collected for P51G (red) and IAIA (blue) variants. The assignments for wildtype CVN are shown for P51G where they could be unambiguously transferred. The spectrum of IAIA is significantly different from that of P51G, but is similarly well-resolved. Certain spectral regions show similar resonance patterns for both proteins (*i.e.* N53–G96–G45–G28–T61–T7–T57 "chain" of resonances at 105–110 ppm in the nitrogen dimension). The resonance assignments could not be transferred onto the spectrum of IAIA, due to a large number of resonances shifting. While the two proteins differ at only three positions, the designed substitutions are in the core of the protein, and their effects (as evidenced by chemical shift perturbation) propagate outward to the surface.

important observation, since it directly demonstrates that a lengthy equilibration time is required to obtain accurate unfolding profiles. The data collected with the shorter overnight equilibration could be forcibly fit to a single unfolding transition, but would yield an apparent stability significantly higher than the actual value. We believe that these effects may have complicated prior analyses of CVN thermodynamics.

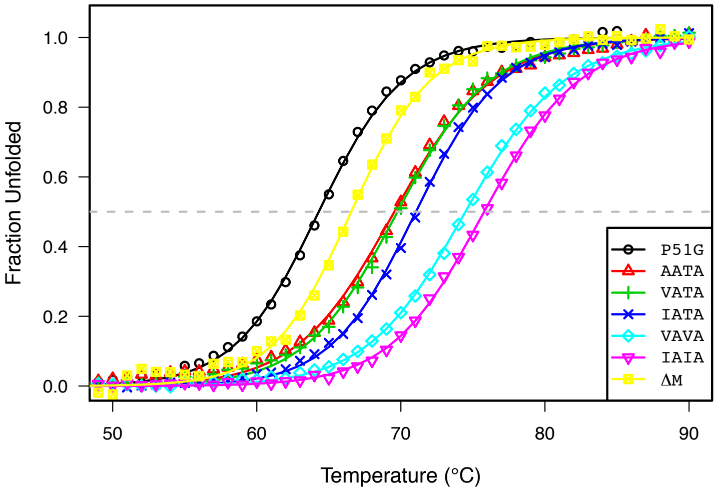


Figure 2-5: Thermal denaturation of CVN mutants followed by CD spectroscopy at 200 nm. The raw data were converted to fraction unfolded; for clarity, only the transition region between 50 °C and 90 °C is shown. The solid lines are nonlinear least-squares fits to the data.

In order to determine the incubation time needed for complete equilibration, we investigated the denaturation kinetics of P51G by following intrinsic tryptophan fluorescence after addition of 3 M GuaHCl (Figure 2-6B). At this denaturant concentration, the protein appears less than 10% folded at equilibrium, but is 60–70% folded after an overnight incubation at 25 °C, as judged by the fluorescence emission intensity. The recorded denaturation kinetics were complex and could not be modeled by a single exponential decay, or by a sum of two decays, as judged by the non-randomness of the residuals. This implies that P51G denaturation proceeds through at least one kinetic intermediate state, the decay of which is slow. These experiments allowed us to determine the apparent half-life of denaturation (13 h) at 3 M GuaHCl and the time for complete equilibration of CVN variants; 72 h was adequate to completely equilibrate all proteins discussed in this work.

To investigate the equilibrium stability of CVN variants to chemical denaturation, we monitored their intrinsic protein fluorescence and circular dichroism signal. The two techniques allowed us to monitor two probes, with fluorescence monitoring the burial of the unique Trp fluorophore of CVN and CD monitoring the state of the polypeptide backbone. Figure 2-7 shows far-UV circular dichroism spectra of P51G taken at equilibrium as a function of increasing concentrations of GuaHCl. These spectra lack an isodichroic (isosbestic) point, indicating that the equilibrium folding is also not two-state [42], but rather that the equilibrium unfolding of CVN by GuaHCl is

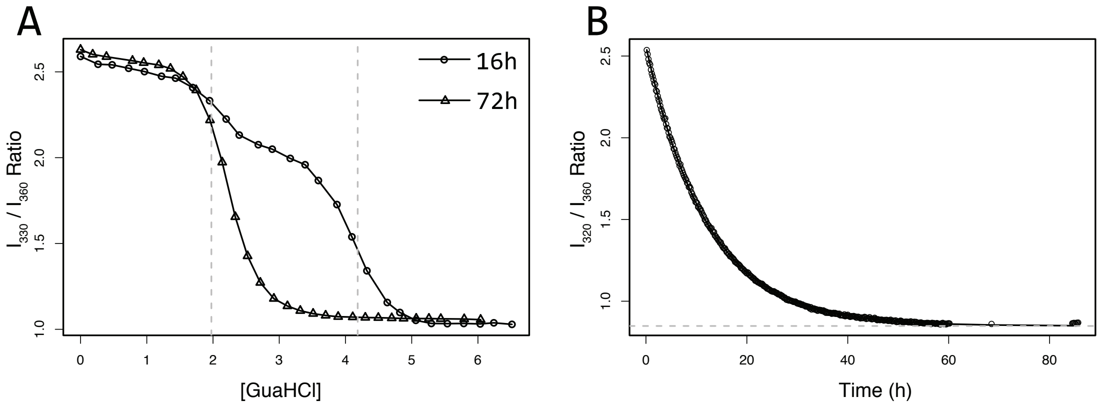


Figure 2-6: **A**. The denaturation profile of P51G monitored by Trp fluorescence emission, after 16- and 72-hour incubations at room temperature in the presence of GuaHCl. The 16-hour profile is biphasic, with dashed grey lines showing the apparent denaturation midpoints of the two phases when they are fitted separately. The numerical derivative of the 16-hour denaturation profile showed two inflection points, at the apparent midpoints of the two phases (not shown). In contrast, 72-hour denaturation profile shows a single transition. These data suggests the presence of a kinetic intermediate in CVN denaturation by GuaHCl. **B**. Fluorescence-monitored denaturation kinetics of P51G after addition of 3 M GuaHCl. A solid line is the fit of the data to a biexponential decay. The apparent half-life of unfolding is 14 h. Note that the data could not be adequately fit by a single or double exponential, as judged by the randomness of the residuals.

complicated, and populates intermediate states. While both kinetic and equilibrium experiments revealed the presence of multiple states, it is unknown whether intermediate states present in the kinetic and equilibrium denaturation of CVN are structurally related.

In order to extract the thermodynamic parameters from the far-UV CD spectra of CVN variants we employed a thermodynamic model which decomposes spectroscopic data into the spectra of three thermodynamic states and their fractional populations, as given by their free energy differences [78, 100, 158]. The data were fit globally at all wavelengths in order to obtain the most reliable parameters. We note that no single far-UV wavelength showed a clear biphasic transition. When the data were fit to a simple two-state model, the derived thermodynamic parameters were dependent on choice of wavelength — for instance, a plot of apparent fraction unfolded at θ_{220} was non-coincident with fraction unfolded at θ_{235} (data not shown). Fluorescence spectroscopy was not sensitive to the presence of an equilibrium intermediate, and thus tryptophan emission spectra were not used in the thermodynamic model.

The parameters derived from the spectral decomposition are summarized in Table 2.4. We found that substitutions at positions 11, 20, and 61 stabilized *both* the native and intermediate states. Substitutions of Ser11 and Ser20 with Ala stabilized the native state by 0.94 kcal mol^{-1} relative to P51G, in agreement with the computational prediction. Further substitutions at position 11 to Val and Ile stabilized the native state by 1.03 and 2.56 kcal mol^{-1}, respectively, relative to P51G.

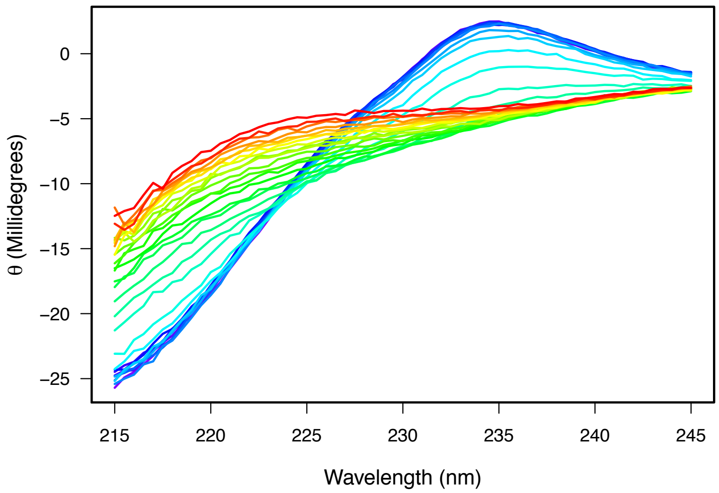


Figure 2-7: Far-UV CD wavelength scans of P51G taken at increasing concentrations of GuaHCl after 72 h equilibration at 25 °C, pH 6.0. The spectra are from 0 M (blue) to 6 M (red) of denaturant. The absence of an isodichroic point at equilibrium demonstrates that the denaturation of CVN by GuaHCl is not strictly two-state.

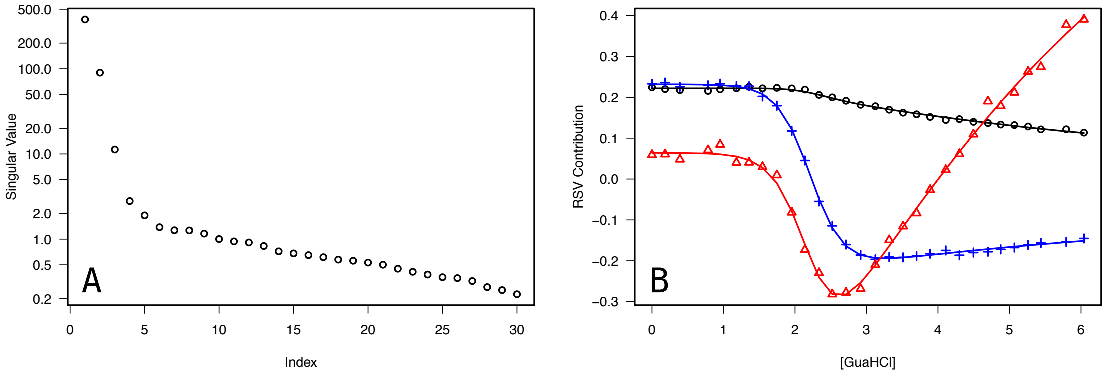


Figure 2-8: **A**. Singular values of GuaHCl-dependent CD spectra shown in Figure 2-7. The algorithm reveals a lower bound on the number of linearly-independent components needed to reconstruct the experimental data. In the case of P51G, three components are present above the noise. The corresponding singular values are $\sigma_1 = 378.93$, $\sigma_2 = 90.07$, and $\sigma_3 = 11.28$. The three-component reconstruction is able to explain 99.98% of the variance in the original data; the variance is partitioned as 94.56%, 5.34%, and 0.08% for the first, second, and third component, respectively. The remaining 27 singular values account for 0.02% of the experimental variance. **B**. GuaHCl-dependent contributions of the right singular vectors. The left singular vectors (not shown) for the three components are strongly autocorrelated: $\text{ACF}(\vec{u}) = \{0.95, 0.98, 0.94\}$. The right singular vectors are similarly autocorrelated: $\text{ACF}(\vec{v}) = \{0.92, 0.94, 0.90\}$. All remaining left and right singular vectors are not strongly autocorrelated, and are deemed random.

Variant	$\Delta G^\circ_{N\to I}$	$m_{N\to I}$	$\Delta G^\circ_{N\to D}$	$m_{N\to D}$	$\Delta\Delta G^\circ_{N\to I}$(P51G)	$\Delta\Delta G^\circ_{I\to D}$(P51G)
P51G	5.55	2.32	6.45	2.70		
AATA	5.56	2.10	7.39	2.64	0.01	0.94
VATA	6.34	2.12	7.48	2.57	0.79	1.03
IATA	7.28	2.27	9.01	2.71	1.73	2.56
VAVA	5.44	1.65	6.75	2.13	−0.11	0.30
IAIA	6.83	1.91	8.49	2.41	1.28	2.04
ΔM	5.80	2.10	6.78	2.41	0.25	0.33

Table 2.4: **Equilibrium denaturation parameters for CVN variants.** The parameters derived from global decomposition of circular dichroism spectra taken as a function of denaturant concentration. The data were fit to a three-state model, as discussed in the text. All free energy values are in kcal mol^{-1}.

Substitutions at position 61 (VAVA and IAIA) were accompanied by significant m-value changes. The m-value for the native to intermediate transition for VAVA is 1.65 while it is 2.32 for the P51G background and 2.1 for ΔM. For VAVA, the free energy difference between the native and denatured states ($\Delta G^\circ_{N\to D}$) is 6.75 kcal mol^{-1}, which is comparable to the stability of the ΔM variant (6.78 kcal mol^{-1}). However, the denaturant concentration at which the VAVA unfolding transition is 50% complete is 0.4 M GuaHCl higher than that of ΔM. This is illustrated in Figure 2-9, which shows the equilibrium denaturation profiles for the designed CVN homologues. The profiles are generated from the equilibrium thermodynamic parameters, and give the fractional populations of the native and intermediate states for all CVN variants as a function of GuaHCl. We found that the equilibrium intermediate state remains significantly populated for all variants at the highest denaturant concentrations tested (up to 6 M GuaHCl), and that all variants populate the intermediate state up to 25% under some experimental conditions.

2.3.7 Carbohydrate Microarray Binding Analysis Demonstrates that Variants Retain Native Binding Specificity

In order to assess the function of our designed variants we carried out fluorescence-detected glycan microarray binding experiments [22] with P51G included as a control. The variants (ΔM not included) were heterogeneously labeled with NHS-Fluorescein at a surface lysine side chain, with the amount of labeling estimated to be approximately 0.5 dye molecules per protein molecule.

Microarray binding experiments confirmed that the designed variants retain the exquisite carbohydrate binding specificity of wild type CVN. The Consortium for Functional Glycomics (CFG) mammalian printed array (version 4.1) contains 465 diverse carbohydrates, featuring both linear and branched molecules. Consistent with previous microarray studies, we observed that the designed mutants bound exclusively to carbohydrates containing the Manα(1→2)Man disaccharide, which has been previously demonstrated to be the minimal carbohydrate recognized by the two independent binding sites of CVN [22, 94]. Figure 2-10A shows the raw fluorescence signal for our

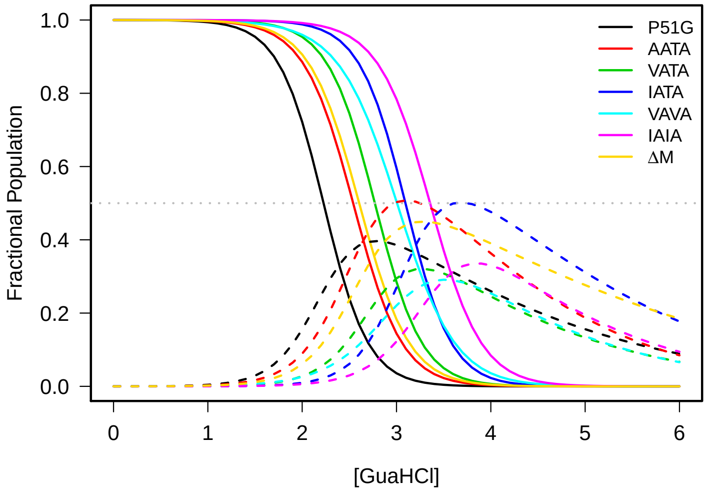


Figure 2-9: Fractional populations of CVN variants derived from decomposition of circular dichroism spectra acquired at increasing concentrations of GuaHCl. Solid lines show the fractional populations of the native state; matching dashed lines show the fractional populations of the equilibrium intermediate state for each variant. The dotted grey line indicates 50% population, and is provided for visual reference. The details of the thermodynamic models used to generate these populations are described in the text.

CVN variants. For all variants, the fluorescence emission followed a decaying profile when ranked by the average intensity across all variants, and we observed binding to 10 distinct carbohydrates, shown in Figure 2-10B. Each of the bound carbohydrates, with the exception of 212, corresponds to a substructure of $\mathrm{Man_9GlcNAc_2}$, the largest (tri-antennary) N-linked high-mannose oligosaccharide present on the viral surface. Compound 212 is equivalent to the D3 arm of $\mathrm{Man_9GlcNAc_2}$ extended by an additional Manα(1→2)Man at the non-reducing end.

As expected, the designed homologues bound to oligosaccharides corresponding to the D1, D2, and D3 arms of $\mathrm{Man_9GlcNAc_2}$; the highest fluorescence was observed for glycan 314, which contains all three arms. The second-highest fluorescence was for glycan 212 and the five largest signals were seen for carbohydrates containing the Manα(1→2)Manα(1→2)Man trisaccharide. This is consistent with the oligosaccharide specificity previously observed in a computational study of CVN–carbohydrate binding [58]. We must note that the fluorescence intensity cannot be interpreted as affinity for a particular carbohydrate, due to possibilities of a nonuniform degree of labeling and potential differences in density and presentation of various carbohydrates on the microarray slides.

Our stabilized CVN variants showed carbohydrate recognition specificity identical to that of P51G. The proteins did not bind to the large majority of the carbohydrates present on the microarray. Two such compounds (49 and 214) represent the core of $\mathrm{Man_9GlcNAc_2}$ and are shown in Figure 2-10. Consistent with previous studies [22, 94], our mutants specifically bound to the terminal arms of $\mathrm{Man_9}$, and not the glycan core.

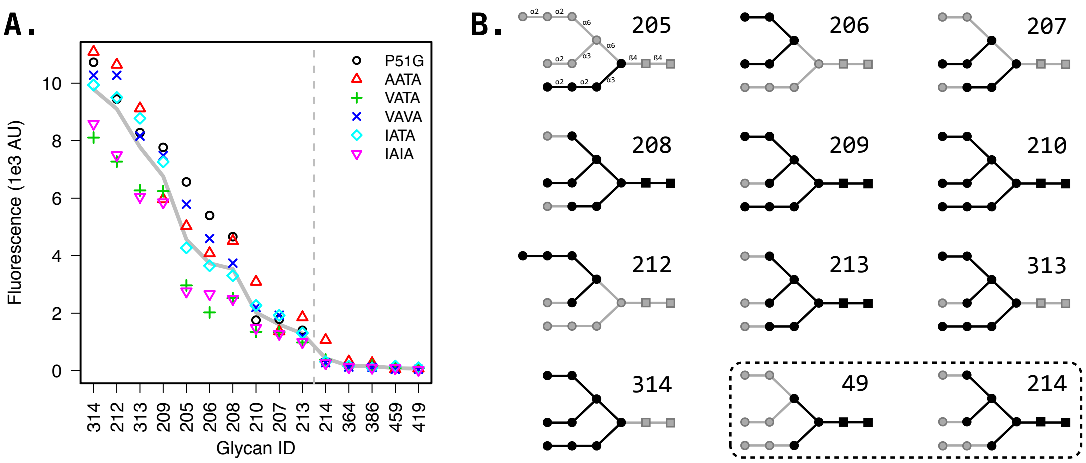


Figure 2-10: **A**. Glycans bound by CVN mutants, ranked by decreasing average fluorescence amongst designed variants. The raw fluorescence intensity of binding to microarray is shown in points. Average fluorescence of all mutants is shown as a thick grey line. Glycan ID corresponds to glycans found on Version 4.1 Glycan Microarray at the Consortium for Functional Glycomics. The dashed vertical line indicates the cutoff for the 10 carbohydrates determined to bind CVN variants. No additional carbohydrates gave rise to signal distinguishable from the experimental noise. **B**. Glycans bound by CVN variants on carbohydrate microarray (shaded in black). For each glycan, the ID is given, as well as the substructure of Man_9 that the glycan corresponds to; Man_9 is shown as background in light gray. Mannose residues are represented by ● and GlcNAc residues by ■. The dashed box contains two examples of Man_9 core carbohydrates which were *not* bound by any of the CVN variants discussed here; these are included to illustrate the specificity of binding to Man$\alpha(1\rightarrow2)$Man disaccharides.

2.4 Discussion & Conclusions

CVN contains a number of polar side chains which are sequestered from solvent. Of the 27 side chains which have a calculated average solvent accessibility below 15%, 10 have polar character. These include the disulfide-bonded Cys8/Cys22 and Cys58/Cys73, Trp49, and Asn42 and Asn93. The latter two residues expose 6% of the Asn surface area, but are highly conserved (Figure 2-1B), and their side chains make key hydrogen bonding interactions which stabilize the two domains of CVN [17]. Poisson–Boltzmann calculations suggest that both of these residues are able to overcome their desolvation penalties of +4.30 kcal mol^{-1} through favorable electrostatic interactions, and that replacing either Asn42 or Asn93 with a hydrophobic isostere is disfavored by +3.30 and +3.01 kcal mol^{-1}, respectively. The remaining three buried polar groups are Ser11, Ser20, and Thr61. Using continuum electrostatic calculations we quantified the desolvation penalty paid by Ser11, Ser20, and Thr61 upon burial within the core of CVN. The single replacement of any of the three groups with a hydrophobic isostere was predicted to be electrostatically unfavorable. However, the Ser11.Ser20 pair was found to contribute unfavorably to CVN stability, and the calculations suggested that the simultaneous replacement of the two serine side chains with hydrophobic isosteres could lead to increased protein stabilization. This analysis highlights the importance of considering multiple mutations during protein design and is similar to observations made with buried salt bridges. The replacement of a single residue in a salt bridge or a polar unit is often highly destabilizing. For instance, the substitution of any charged residue with Ala in a complex "Arg–Glu-Arg" salt bridge triad within Arc repressor has been demonstrated to have a detrimental effect on protein stability and function. However, the simultaneous replacement of the triad with "Met–Tyr–Leu" yielded a variant significantly more stable that wild type [181]. Pairs of buried interacting polar residues can be thought of as analogous to buried salt bridges, albeit with less overall charge density.

In contrast to Ser11 and Ser20, Poisson–Boltzmann calculations suggest that the electrostatic contribution of Thr61 to the stability of CVN is neutral (Table 2.1). However, we found that the protein could be stabilized by substitutions at position 61. For example, VAVA was 4.7 °C more thermostable than VATA, and IAIA was more thermostable than IATA by the same amount. The observed stabilization may arise due to enhanced van der Waals packing interactions, or may indicate an overestimation of intermolecular interactions with respect to the desolvation penalty in the calculations.

In agreement with the sequence analysis of the CVN family, the computational redesign of Domain A successfully predicted that the protein is able to accommodate Ile, Val, or Ala at position 11. The computational redesign of Domain B, where Thr61 and Ala71 were allowed to vary in sequence, predicted that only a single sequence (T61A.A71) to be more favorable than wild-type (by −2.4 kcal mol^{-1}). A number of additional sequences, such as T61V.A71 were predicted to be significantly worse (+10.7 kcal mol^{-1}). This could indicate that Domain B is less tolerant of repacking or could highlight limitations of the fixed-backbone model. In this case, since experimental characterization demonstrated stabilization with substitutions at Thr61, the result is most likely an artifact of the fixed-backbone approximation, an effect which has been observed previously [73, 120].

The thermostabilization of our designed CVN variants relative to P51G most likely includes contributions from several factors. The initial 5.3 °C stabilization of AATA relative to P51G likely arises primarily from the removal of the unsatisfied polar groups at positions 11 and 20. Alanine is smaller than serine in volume, so the substitutions are unlikely to add van der Waals packing interactions. Intriguingly, VATA does not have significantly improved thermostability relative to AATA, despite the increased hydrophobicity and size of the Val residue compared to Ala. The IATA mutant has the largest side chain that we expected could be successfully incorporated at position 11, and it showed improved thermostability relative to AATA by 1.5 °C.

We found that the stabilization of the designed variants can be qualitatively explained by considering the additional nonpolar surface area introduced by the mutations [46]. Table 2.5 lists nonpolar surface areas calculated for all the mutants at positions 11, 20, 61, and 71, as well as the free energy difference ($\Delta\Delta G^{\circ}{}_{\phi}$) expected from the surface area change; the $\Delta\Delta G^{\circ}{}_{\phi}$ is estimated using an "atomic solvation parameter" of $16\,\text{cal}\,\text{Å}^{-2}\,\text{mol}^{-1}$ [46].

Variant	SASA_{np} (Å^2)	ΔSASA (Å^2)	$\Delta\Delta G^{\circ}{}_{\phi}$ ($\text{kcal}\,\text{mol}^{-1}$)
P51G	238	—	0.00
AATA	283	45	0.72
VATA	328	90	1.44
IATA	351	113	1.82
VAVA	367	129	2.07
IAIA	415	177	2.83

Table 2.5: **Predicted thermodynamic stabilization based on buried nonpolar surface area.** Nonpolar surface area was calculated for the side chains at positions 11, 20, 61, and 71 using standard parameters provided with NACCESS. ΔSASA is the estimated change in nonpolar surface area upon mutation, relative to P51G. The expected change in the transfer free energy ($\Delta\Delta G^{\circ}$) was calculated according to Eisenberg [46].

Our kinetic denaturation studies of P51G reveal that at room temperature the protein requires nearly three days to fully reach equilibrium at intermediate GuaHCl concentrations and suggest that at least one kinetic unfolding intermediate is populated. Equilibrium circular dichroism spectra taken at increasing concentrations of GuaHCl showed an absence of an isodichroic point, and are also incompatible with a two-state model of folding. Singular value decomposition [78] of GuaHCl-dependent CD spectra revealed the presence of three significant spectral components above the experimental noise (Figure 2-8), consistent with multistate equilibrium denaturation.

A number of previous studies [9, 10, 126] have examined the equilibrium chemical denaturation of wild-type CVN, P51G, and several functional homologues using intrinsic protein fluorescence. Those studies reached a different conclusion about the stability of CVN, and in particular, the denaturation of CVN was deemed two-state. We believe that the present work provides more precise thermodynamic parameters and apparent stabilities than were previously obtained, since earlier studies measured protein stability after overnight incubation with GuaHCl. Overnight incubation is insufficient to allow the system to equilibrate, and this leads to estimated stabilities which are larger than the actual thermodynamic stability [10, 126].

The structural reason for the slow denaturation of the P51G variant of CVN is unknown. While this variant does not contain any proline residues, its disulfide bonds may contribute to the slow denaturation kinetics, as cystine residues have strong preferences for specific C–S–S–C dihedral angles and high energetic penalties for deviation from them [74]. For example, both CHARMM [115] and AMBER [36] molecular mechanics force fields implement this dihedral angle as two isoenergetic minima ($\chi_{\mathrm{SS}} = \pm 84°$) separated by a barrier height of $6.4\,\mathrm{kcal\,mol^{-1}}$, comparable to that of the rotation about the peptide bond. Similarly slow unfolding kinetics have been observed for disulfide-containing hen egg white lysozyme [101]. However, the slow unfolding kinetics may also be due to other structural factors, having also been observed for the cystine-less four-helix bundle Rop [130].

We found that the presence of an N-terminal methionine residue has a measurable effect on CVN stability. The ΔM variant possessed increased stability when compared to P51G (Table 2.3). Biophysical studies of the B1 domain of staphylococcal protein G (β1) reported a similar effect; the methionine-containing and methionine-lacking forms of the protein differ by $1.7\,\mathrm{kcal\,mol^{-1}}$ in stability [167]. Similar effects have been observed for α-lactalbumin [32, 33]. For CVN, the effect is not as drastic, yet is significant; we hypothesize that the difference between the stability of P51G and ΔM stems from the fact that the *native* N-terminus, which has the methionine removed, participates in favorable electrostatic interactions in the folded state.

Chapter 3

Characterization of the Individual Domains of the Lectin MVL and Design of Stabilizing and Monomerizing Mutations[*]

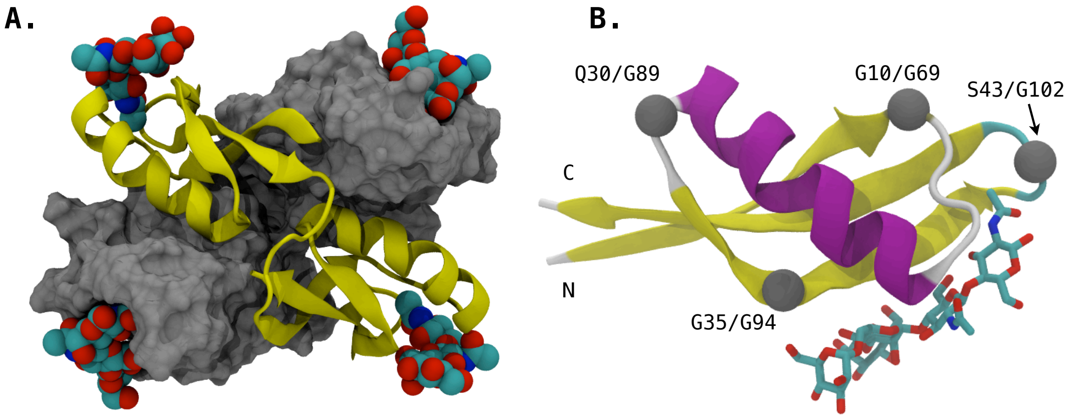
Abstract

Numerous carbohydrate-binding proteins inhibit infection by HIV by binding to the glycoprotein gp120 on the surface of the virus. This glycoprotein contains approximately two-dozen putative glycan targets, not all equally accessible to inhibitors. While carbohydrate-binding proteins are commonly multivalent, they are *uni*specific, and recognize only several of the many putative glycan targets. Simultaneously targeting several distinct glycan epitopes may yield inhibitors beyond the potency of natural compounds. Furthermore, the need for multivalency for potent HIV neutralization is debated. In this work, we present efforts to engineer a combinatorial carbohydrate recognition scaffold against the diverse glycan targets on gp120, based on the HIV-inhibitory lectin MVL. We characterized the isolated domains of MVL, and discovered that while the N-terminal domain is unfolded under all conditions, the C-terminal domain was folded and dimeric. By testing the inhibitory potency of the C-terminal domain of MVL side-by-side with the tetravalent protein, we revealed that the two possess near-identical IC_{50} values, arguing that multivalency is not necessary for MVL action against HIV.

3.1 Introduction

In this work, we biophysically characterized individual domains of the lectin MVL [16,160,188,196]. MVL (*Microcystis viridis* lectin, UniProt ID: `Q9RHG4`) was originally isolated due to its ability to agglutinate rabbit erythrocytes. While the function of the protein within its native organism is presently not known, MVL inhibits infection by the Human Immunodeficiency Virus

[*] The *in vitro* HIV neutralization experiments presented in this chapter were done in collaboration with the laboratory of Dr. Carole Bewley, and are included here with permission.



C.

```
                        10        20        30          40        50
Q9RHG4/N       MASYKVNIPAGPLWSNAEAQQVGPKIAAAHQ--GNFTGQWTTVVESAMSVVEVELQVENTGI
Q9RHG4/C       -HEFKTDVLAGPLWSNDEAQKLGPQIAASYG--AEFTGQWRTIVEGVMSVIQIKYTF-----
                         70        80          90       100       110
A9DK39/N       MGKFTVSIPAGPIWNDEDGKEKGPIVAAAHL--GEFDGNWRTVVPNEMSTVDVILNSEPTGS
A9DK39/C       -SEYTLDVLAGPIWNQEDAEKKCPVVCASYG--GKWNGQWKTVVSGKMSVCGCTFKF-----
A8YAZ8/N       SNEEPVNIPAGPLWSNAEAQQLDPRIAAAHQ--GNFTGQWTTVVESAMSVVEVELQVENTGT
A8YAZ8/C       -HEFKTDVLAGPLWSNDEAQKLCPQIAASYG--AEFTGQWRTIVEGVMSVIQIKYTF-----
Q28L60/N !     AASAQQAFDAGPIWDQNHANQVCPAVAASHG--GTWTGHWWTTVPNQMSVCQVQVASPA---
Q28L60/C !     -----IAVEAGPIWNQNHANQVCPRLAASIR--GTWTGQWWTTQPSVMSVCQIIP-------
A9G5N8/N !     -------LEAGPIWNTSDAQTKCPNVCNPQN--MSWNGQWWTTVPGAMSVCECAPRPAAV--
A9G5N8/M !     -------VQAGPIWSNTHAQTQCPNTCAAYSSATKWNGQWWTTVPGQMSVCECAFTPPAT--
A9G5N8/C !     -----VSLEAGPIWSNADAPSKCPAACGTSR---AWNGQWSTSVAGQMSVCGCACTP-----
A0NYB3/N !     AGAQTYNVEAGPIWNNGDAQAKCPRVCGGLG--TRWNGQWHTTVQGQMSVCSCEKASPG---
A0NYB3/C !     -----RDIDAGPIWNNADAQGKCPGICFGNG--LSWSGQWRTTVQGRMSVCECR--------
A9NDM7/C !     HYRFVKDFPAGPIWNQADAQNKCPPVCASHG--ARWTGNWHTVREGRQSVCQCRGWSRWSR-
A9ZJJ6/C !     HYRFVKDFPAGPIWNQADAQNKCPPVCASHG--ARWTGNWHTVREGRQSVCQCRDWSRWSR-
                      . ***:*.  ..    *  .         : *:* *   .  *.
```

Figure 3-1: **The algal lectin MVL. A**. The MVL homodimer: one chain is shown in surface representation, the second as a ribbon. Each monomer contains two homologous carbohydrate-binding domains, shown in complex with Man$_1$GlcNAc$_2$. **B**. Each domain of MVL is composed of an antiparallel three-stranded β-sheet and a single α-helix. Residues which adopt a positive backbone ϕ angle are indicated by a grey sphere. **C**. A curated alignment of MVL domains derived from Pfam family PF12151 [54]. MVL homologs are composed of two or three tandemly-repeated domains. Each domain sequence is identified by its UniProt accession code, and whether it is an N-, middle, or C-terminal domain of the full polypeptide (/N, /M, or /C, respectively). Domain sequences of MVL (Q9RHG4), along with sequence indices, are shown at the top. Regions of high conservation are shaded grey. The presence of a localization/export tag at the N-terminus of the hypothetical polypeptides is indicated by "!". This tag correlates with the cysteine residues at alignment positions 22, 28, 48, and 50 (highlighted in blue); structural modeling suggests that these form disulfide bonds.

(HIV) by binding gp120 oligosaccharides on its surface, and has been shown to recognize the Man$\alpha(1 \rightarrow 6)$Man$\beta(1 \rightarrow 4)$GlcNAc$\beta(1 \rightarrow 4)$GlcNAc tetrasaccharide core common to both high-mannose and complex N-linked glycans. MVL is a 12 kDa, 113-residue protein, containing two 54-residue domains separated by a short linker. Biophysical studies, leading ultimately to the determination of the X-ray structure of MVL, revealed that the protein is an obligate homodimer with C2-like pseudosymmetry (Figure 3-1A). MVL has four carbohydrate binding sites, one in each domain, which are of identical affinity and specificity for their cognate ligands. Furthermore, the sequence identity of the binding sites in the two domains is almost perfectly conserved. Structural details of carbohydrate recognition by MVL are discussed in the subsequent Chapter 4. First, we aim to determine whether MVL is a good starting point to engineer novel carbohydrate specificity, in order to screen various carbohydrate specificity combinations against a diverse target. Second, we aim to determine whether multivalency is required for potent neutralization of HIV by MVL.

Lectins are typically uni-specific. Lectins often contain modular carbohydrate-recognition domains, and can be composed of single polypeptide chains with tandemly-repeated domains (e.g. actinohivin or Cyanovirin, discussed in Chapter 2), as a noncovalent complex of identical domains (e.g. ConA or pentameric ring domain of cholera toxin), or as a complex of multi-domain polypeptide chains (e.g. MVL) [110]. Despite their multivalency, most lectins recognize a single carbohydrate or a family of closely-related carbohydrates, with few exceptions. The carbohydrate targets on HIV are diverse in their composition and structure. While some lectins inhibit viral infection by binding to the terminal arms of high-mannose oligosaccharides containing $\alpha(1{\rightarrow}2)$-linked mannose residues, others recognize the GlcNAc core, or the $\alpha(1{\rightarrow}6)$ mannose-containing "intermediate" region of Man$_9$GlcNAc$_2$. Gp120 surface glycans are unequal in number, and likely in their accessibility and dynamical content.

The role of multivalency in the context of HIV inhibition by lectins is debated. A number of questions remain about the mechanism by which MVL inhibits HIV, one of the most pressing of which is whether protein multivalency is necessary for potent viral neutralization. Unlike CVN, which binds carbohydrates with µM affinity yet is a single-digit nM inhibitor, MVL binds its cognate sugars and has a concentration of 50% inhibition (IC$_{50}$) in the low nM range [16]. This suggests that the tetravalency of MVL does not lead to significant cooperative affinity enhancement toward gp120. Mannose-recognizing lectins bind high-mannose sugars at their termini (D1, D2, and D3 arms, Figure 4-1B), which are furthest from the glycoprotein and are the most solvent-exposed. MVL, on the other hand, recognizes the portion of Man$_9$ closest to the glycoprotein, which is likely the most rigid. Carbohydrate recognition by MVL requires geometric/orientational specificity beyond that of mannose-recognizing lectins (discussed in Chapter 4). Consequently, it may be more difficult for multiple MVL binding sites to simultaneously engage the sugars on a single gp120 trimer, or to bridge across trimers.

The vast majority of proteins which inhibit viral infection at nanomolar affinity or below are multivalent. Mutations which abolish either of the two binding sites of the bivalent CVN severely abrogate its inhibition, and variants featuring more binding sites by virtue of dimerization or tandem duplication have enhanced potency [56, 90, 92]. A recent effort to create a monomeric variant of the lectin Griffithsin also revealed the need for multivalent interactions in order for this lectin to inhibit infection at sub-µM concentrations [129]. Microvirin, a close homologue of cyanorivin-N, represents an intriguing counter-example: the protein contains a single carbohydrate binding site, but is a potent HIV inhibitor. Since CVN and microvirin recognize the same Man$\alpha(1 \rightarrow 2)$Man target present at the terminal arms of the high-mannose saccharide Man$_9$ with similar affinities, at the present time it is unclear how this monovalent lectin functions [18, 161]. In general, the mechanism of HIV inhibition by lectins is poorly understood. In the case of CVN, where the presence of both binding sites is essential for potent inhibition, it has been hypothesized the protein may function by agglutinating viral particles or by cross-linking gp120 glycans [48]. HIV neutralization by lectins or antibodies has a time-dependence, further complicating mechanistic investigations [161].

3.2 Design of a Multivalent Lectin Scaffold for Combinatorial Screening against Polyvalent Glyco-ligands

Tools are needed in order to explore the role of multivalency in the context of HIV inhibition. Monovalent lectins (microvirin or engineered CVN mutants) have begun to address these questions in the context of binding to $\alpha(1 \rightarrow 2)$-linked mannose epitopes [9, 31, 161, 163]. In addition, designed lectins with different specificity for HIV carbohydrate epitopes in their binding sites can, in principle, be as potent as *uni*specific lectins studied presently. A scaffold which facilitates combinatorial screening of carbohydrate binding specificities against a diverse target such as gp120 may answer questions about distance and geometry of the various carbohydrate epitopes on the glycoprotein (Figure 3-2). We set out to determine whether the tetravalent lectin MVL can be used as a starting point for scaffold design.

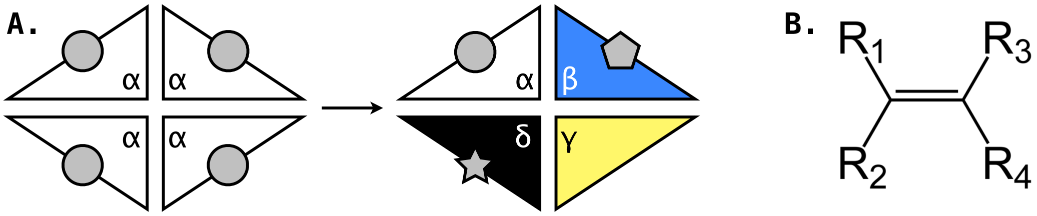


Figure 3-2: **Design of a "mix-and-match" carbohydrate-binding scaffold. A.** Starting from a tetravalent protein composed of identical subunits α specific for the same carbohydrate ligand (circle), we aim to design a hetero-tetrameric scaffold composed of distinct subunits $(\alpha, \beta, \gamma, \delta)$ with orthogonal binding specificities (or null affinity for "knockout" studies.) Such a molecule would facilitate the exploration of multivalency in carbohydrate binding, aid in profiling local glycosurface geometry, and allow combinatorial screening against different carbohydrate epitopes. The schematic is based on MVL. **B.** The combinatorial complexity of the scaffold shown in Panel **A** is the same as that of functional derivatization of a tetrafunctional alkene.

Combinatorial Complexity of Tetramer Assembly. The assembly of multimers out of homologous building blocks faces the problem of specificity. In particular, multimeric assembly proceeds through a large number of intermediate states, many of which lead to non-productive ("off-pathway") complexes. In the case of a tetrameric scaffold based on the individual domains of MVL, we enumerated the putative assembly intermediates and tetrameric products using two building blocks (Figure 3-3A). The combinatorial complexity of multimeric assembly increases exponentially with the number of building blocks. For instance, four monomers can produce 76 distinct tetrameric states with MVL-like pseudosymmetry, six of which are non-identical $\alpha\beta\gamma\delta$ heterotetramers (Figure 3-3B).

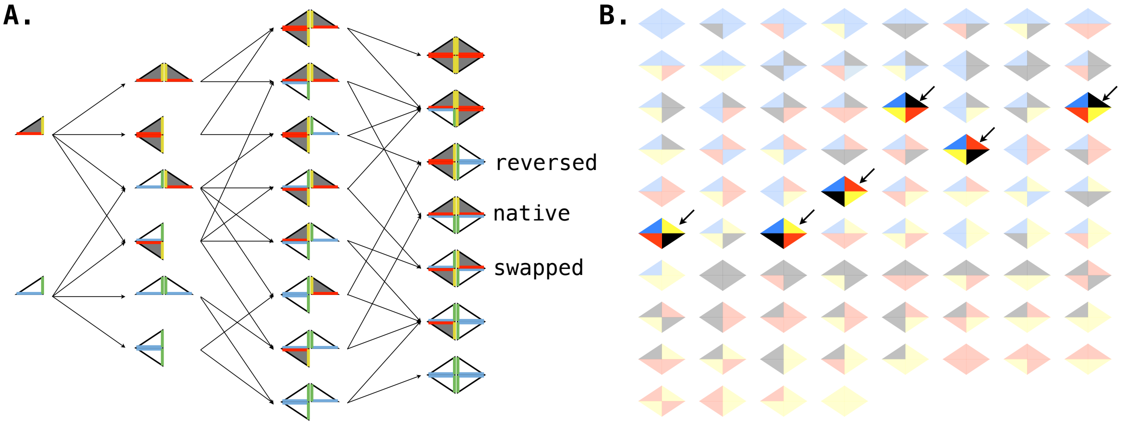


Figure 3-3: **Combinatorial complexity of tetramer assembly. A**. Assembly of tetramers starting from monomeric building blocks. The N- and C-terminal domains of MVL can be schematically represented as a scalene right triangle (grey and white, respectively.) Each domain contains two non-overlapping association interfaces (colored red, yellow, green, and blue) and a carbohydrate-binding interface (hypotenuse, black). Assembly of the seven possible tetramers proceeding through all possible intermediates is shown. Three of the fully-assembled tetramers contain two N- and C-terminal domains each, and are labeled. The "native" $\alpha_2\beta_2$ complex employs the interfaces used in wildtype full-length MVL. The "reversed" complex corresponds to full-length MVL in which the orientation between the two chains is flipped. The "swapped" complex corresponds to a heterodimer of full-length MVL, with one chain composed of two $\mathrm{MVL_N}$ domains and the other of two $\mathrm{MVL_C}$ domains. Figure inspired by Williamson [189]. **B.** Possible heterotetramer assemblies using four distinct building blocks. The complexity of tetramer assembly based on MVL as the scaffold is equivalent to counting the number of ways of coloring the sides of a rhombus (symmetric about both axes) using n colors, allowing turning over (also see Figure 3-2B). For two building blocks, $A(2) = 7$ (Panel **A**); in the case of four building blocks, $A(4) = 76$ possibilities arise, six of which are non-identical $\alpha\beta\gamma\delta$ heterotetramers (indicated with arrows). The seventy nonproductive tetramers which contain multiple copies of any particular building block are shown with reduced opacity. The high number of nonproductive states underscores the complexity of the design problem.

3.3 Materials & Methods

Cloning, Protein Expression & Purification. The plasmid encoding wildtype MVL was generously provided by Dr. Carole Bewley (National Institutes of Health). The full-length gene, or portions corresponding to residues 1–54 or 60–113 were PCR-amplified with appropriate primers, and subcloned into the pE-SUMO vector (LifeSensors) using BsaI and XbaI restriction sites; this plasmid encodes the gene for yeast Smt3 SUMO1 homologue with an N-terminal polyhistidine tag.

To produce recombinant MVL variants, appropriate plasmids were transformed into BL21(DE3) strain of *E. coli*; cells were grown in LB medium supplemented with kanamycin sulfate (30 µg mL^{-1}) at 37 °C with shaking until the optical density reached 0.8. Protein expression was induced with 1 mM IPTG and was continued overnight at 20 °C. Cells were collected by centrifugation and stored at −80 °C until purification. ^{15}N-labelled proteins were produced according to standard protocols [79] with ^{15}N-enriched ammonium chloride (Cambridge Isotopes) as the only source of nitrogen.

Cells were lysed by sonication, and recombinant proteins were purified under native conditions using a 5 mL HisTrap affinity column (GE Healthcare) with an imidazole gradient (up to 0.3 M). Fractions containing SUMO fusion proteins were treated with 200 µg of Ulp1 SUMO protease, and dialyzed overnight against protease cleavage buffer (20 mM Na-phosphate, 150 mM NaCl, 0.5 mM dithiothreatol, pH 7.5) at 4 °C. After cleavage the reaction was passed through the HisTrap column equilibrated with cleavage buffer to capture the uncleaved proteins, and His-tagged Smt3 and Ulp1 proteins; finally, the proteins were purified using size exclusion chromatography (Superdex 75 26/60) in buffer A (20 mM Na-phosphate, pH 7.5) or in PBS.

Hydrodynamic Characterization. Analytical gel filtration was performed on a Superdex 75 (10/300 GL) column (GE Healthcare). Samples were injected onto the column pre-equilibrated in 20 mM sodium phosphate, 200 mM sodium chloride (pH 7.5) and were isocratically eluted at 0.5 mL min^{-1}. Sedimentation equilibrium analytical ultracentrifugation experiments were typically conducted at 25 °C on a Beckman Optima XL-A analytical ultracentrifuge at three rotor speeds. The solute partial specific volumes were calculated based on protein amino acid composition, and the solvent density was estimated from standard tables. Data were fit globally using HeteroAnalysis (University of Connecticut Analytical Ultracentrifugation Facility).

Protein Stability Determination. Samples for equilibrium denaturation studies were prepared as follows. Typically, thirty-two 2 mL samples containing 10 µM protein in buffer A with an appropriate concentration of GuaHCl were incubated at room temperature for 24 h. The concentration of GuaHCl in each sample was determined by refractometry. Samples were analyzed by circular dichroism spectroscopy by following ellipticity at 226 nm.

Molecular Dynamics Simulations. All system setup and post-processing steps were performed with CHARMM [27]. Individual MVL domains were solvated in an orthorhombic box contain-

Plasmid	Gene	Mutation(s)	Result
pESM01	MVL		Folded protein
pESM02	MVL	A16D	5-fold reduced expression
pESM03	MVL	S43G	Unknown
pESM04	MVL	E42P,S43G	Unknown
pESM05	MVL	N6C	Unknown
pESM06	MVL	TEV	No expression
pESMN01	$\mathrm{MVL_N}$		Unfolded protein
pESMN02	$\mathrm{MVL_N}$	S43G	Unfolded protein
pESMN03	$\mathrm{MVL_N}$	Q30G,G31A	Unfolded protein
pESMN04	$\mathrm{MVL_N}$	Q30G,G31A,S43G	Unfolded protein
pESMN05	$\mathrm{MVL_N}$	E42P,S43G	Unfolded protein
pESMN06	$\mathrm{MVL_N}$	Q30G,G31A,E42P,S43G	Unfolded protein
pESMN07	$\mathrm{MVL_N}$	Y3F,V5T,P8A	Unfolded protein
pESMN08	$\mathrm{MVL_N}$	Y3F,V5T,P8A,Q30G,G31A,E42P,S43G	Unfolded protein
pESMC01	$\mathrm{MVL_C}$		Folded protein
pESMC02	$\mathrm{MVL_C}$	L80K	Poor expression
pESMC03	$\mathrm{MVL_C}$	T64D	Poor expression
pESMC04	$\mathrm{MVL_C}$	T64D,L67D	Poor expression

Table 3.1: **List of MVL plasmid constructs made during the course of this study.** All plasmids were constructed starting with pE-SUMO(Kan) (Lifesensors, Inc.)

ing approximately 3300 TIP3P water molecules (4900 for $\mathrm{MVL_C}$ dimer simulations), and charge-neutralized by adding $\mathrm{K^+}$ and $\mathrm{Cl^-}$ ions to a final concentration of 150 mM. To avoid self-interactions in the periodic cell, water box dimensions were set such that the solute was 10 Å away from the cell side. Solvation and neutralization steps used a locally-modified input script originally retrieved from www.charmmtutorial.org. In rare cases when visual inspection identified water molecules inside the protein, the offending molecules were translated into the bulk solvent. The CHARMM22-CMAP [115] molecular mechanics force fields was used.

Isothermal–isobaric (NPT) ensemble molecular dynamic simulations were carried out at 300 K with the NAMD 2.6 engine compiled for the IBM Blue Gene/L architecture [143]. Constant pressure of 1 atm was maintained with the Nosé–Hoover Langevin piston pressure control. All bonds involving hydrogen atoms were constrained to their equilibrium lengths using the SHAKE algorithm. The simulation time step was set to 2 fs and non-bonded interactions were evaluated every step. We used a cutoff of 12 Å for all Lennard–Jones and short-range electrostatic interactions. Long-range electrostatics were treated using a Particle Mesh Ewald method with a fourth-order interpolation scheme. The solvated and neutralized system was minimized for 5000 steps and heated to the production temperature over the course of 200 ps. No restraints were employed during the equilibration or production steps.

Binding Free Energy Calculation. Binding free energies were computed with the MM/PBSA model [96,172]. The total binding energy for each snapshot was computed as the sum of a Poisson–Boltzmann-based electrostatic contribution ($\Delta G^{\circ}_{\mathrm{elec}}$), the intermolecular van der Waals energy ($\Delta G^{\circ}_{\mathrm{vdW}}$), and a term proportional to the solvent accessible surface area buried on binding. In the latter case, the area was computed with CHARMM using a 1.4 Å radius probe and the energetic contribution was given by $\Delta G^{\circ}_{\mathrm{h}\phi} = 0.005 \cdot \Delta \mathrm{SASA} + \ 0.86\,\mathrm{kcal\,mol^{-1}}$ for each snapshot. Binding energies were calculated using snapshots extracted from the solvent-stripped MD trajectories at an interval of 1 ns, as described previously [57].

Spectroscopic Characterization. NMR spectra were recorded at 293 K on a Varian Inova 600 MHz spectrometer using a WaterGate-HSQC pulse sequence. Samples of ^{15}N-labeled MVL or $\mathrm{MVL_C}$ protein was concentrated to 200 µM by ultrafiltration and D_2O was added to 10%. The spectra were visualized using SPARKY (T.D. Goddard and D.G. Kneller, SPARKY 3, University of California, San Francisco).

Circular dichroism spectra were collected using a Chiroscan spectrometer (Applied Photophysics). Far-ultraviolet spectra were collected from 190 nm–260 nm using a 1 mm path-length cuvette. Typically, three spectra were collected at 1 nm increments at a rate of 2 s nm^{-1}, averaged and smoothed using a Savitzky–Golay filter to yield the final result.

HIV Neutralization Assays. Env-pseudotyped HIV neutralization assays were performed as described previously [161] using viral particles pseudotyped with HIV-1 envelope proteins from four different strains. Five-fold serial dilutions of inhibitor were added to the pseudovirus, followed by TZM-bl (CXCR4- and CCR5-expressing cells) target cells at 37 °C. 48 h post-infection, cells were lysed, and luciferase activity was measured. Positive and negative controls were used to normalize the data, and the resulting isotherms were fit to obtain the $\mathrm{IC_{50}}$.

3.4 Results

We aim to determine whether MVL is a good starting point for lectin scaffold engineering. As a first step, we implemented a protein expression system to produce wildtype MVL and its individual domains. We next used this system to express recombinant proteins, and characterized the resulting proteins using spectroscopic techniques.

3.4.1 MVL is Correctly Folded When Expressed as a SUMO Fusion

The SUMO expression system is typically used to prepare proteins which are unstable, or which may be insoluble. In addition to the solubility-enhancing effects of SUMO, one of the great advantages of this system is its ability to produce proteins without cloning artifacts. Previous studies expressed MVL without a purification tag, and purified the protein by fractional ammonium sulfate precipitation and several additional chromatographic steps [16]. Expression of SUMO-MVL fusion protein, followed by Ulp1 protease cleavage, produced correctly-folded MVL. The heteronuclear single quantum coherence (HSQC) NMR spectrum was well-resolved in both the proton and nitrogen dimensions (Figure 3-4A), and showed near-perfect overlap with the previously-determined resonance assignments (provided by C. Bewley) [16,160]. Analytical ultracentrifugation equilibrium experiments revealed that the protein behaves as an ideal solute at all speeds and concentrations. Global analysis of the AUC data sets yielded an average MW of 28.3 kDa, 16% larger than is expected for a homodimer of 12.2 kDa subunits. The discrepancy between the calculated and expected MW for MVL in solution is attributed to instrument mis-calibration; analytical gel filtration experiments demonstrated that the protein elutes as a single symmetric peak at all concentrations. We confirmed the dimeric quaternary structure of MVL by performing concentration-dependent chemical denaturation studies measurements at three protein concentrations (2, 5, and 20 µM). The results, shown in Figure 3-4B, demonstrate that protein stability is concentration dependent.

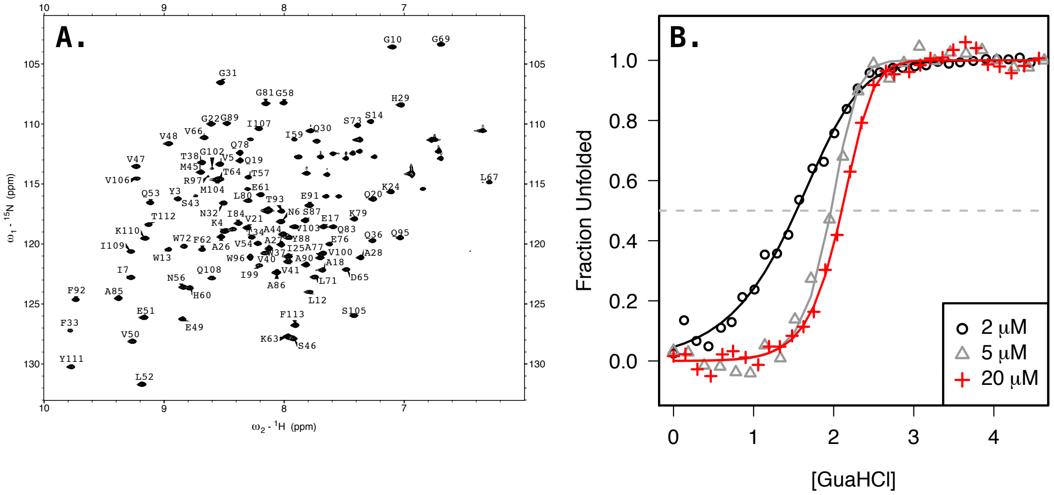


Figure 3-4: **A**. MVL is folded correctly when produced with the SUMO expression system, as judged by the agreement with previously-determined HSQC resonance assignments. **B.** The stability of MVL to chemical denaturation by the chaotrope GuaHCl is concentration-dependent, with higher concentrations requiring greater amounts of denaturant to unfold. These results demonstrate that the protein is a homodimer in solution.

3.4.2 The N- and C-terminal Domains of MVL Differ in Stability

We next expressed SUMO-MVL$_N$ and SUMO-MVL$_C$ independently. At the end of the chromatographic purification sequence, under both native and denaturing purification conditions, we observed that MVL$_N$ was unfolded at 4 °C. In contrast, the MVL$_C$ was folded, and produced a CD spectrum super-imposable over that of wildtype MVL (Figure 3-5A).

Why is $\mathbf{MVL_N}$ unfolded? The sequences of MVL_N and MVL_C differ in several key aspects. First, the MVL domain contains four positions with a positive backbone ϕ angle: G10/G69, Q30/G89, G45/G94, and S43/G102 (sequence indices correspond to the N- and C-terminal domains, respectively). Glycine is the only naturally-occurring amino acid which can adopt a positive backbone ϕ angle without any strain or significant clash of the sidechain [3]. However, two of the four positive-ϕ angle positions in MVL_N are not glycine. Of these, Q30/G89 is a helix-capping position, and S43/G102 is in a type II β-turn. MVL_N also contains a proline residue (P8) in the first β-strand (MVL_C has L67); proline is a known secondary structure breaker. We set out to stabilize MVL_N using the principles of rational design, and created symmetry-restoring variants S43G, Q30G_G31A, and Q30G_G31A_S43G. However, none of these mutants were folded. Type II β turns have strong sequence preference at the $i + 1$ (E42/E101) and $i + 2$ (S43/G102) turn positions, of proline and glycine, respectively [86]. Based on the observed β-turn sequence preference we created variants E42P_S43G and Q30G_G31A_E42P_S43G. However, these variants were also unstructured.

Computational insights into the stability differences between the two domains of MVL. We turned to a computational approach in order to further understand the differences between MVL_N and MVL_C. For each domain, we generated molecular dynamics trajectories of 400 ns in length. These MD trajectories are of modest length, and no unfolding was expected or observed on these timescales. The backbone fluctuations of the two domains in simulation were very similar, with both experiencing fraying of their termini. To understand the differences in the folded-state interactions made in the two domains, we energetically decomposed the resulting snapshots into per-residue van der Waals and electrostatic contributions to protein stability. After calculating the pairwise van der Waals interactions between all protein groups (backbone, amino, and side chain) of the individual domains we found that the sum total of the vdW contributions to MVL_C stability was $16.8\,\mathrm{kcal\,mol^{-1}}$ more favorable than that of MVL_N, indicating more favorable packing of the C-terminal domain. The majority of this difference is a consequence of more favorable side-chain packing interactions of MVL_C.

We calculated the electrostatic contribution of protein sidechains to domain stability using a linearized Poisson–Boltzmann approach. Both domains feature stabilizing salt bridge interactions, most favorable of which are K4/E49 ($-10.3\,\mathrm{kcal\,mol^{-1}}$) and D65/R97 ($-14.4\,\mathrm{kcal\,mol^{-1}}$). MVL_C, however, contains more groups which make more significant contributions (R97, K110, D61, and D65 have $\Delta G^\circ_{\mathrm{mut}}$ values of $-9.7, -6.8, -5.9$, and $-4.8\,\mathrm{kcal\,mol^{-1}}$, respectively); MVL_N most significant contributors are E49 and K4, which contribute -7.9 and $-7.5\,\mathrm{kcal\,mol^{-1}}$ to domain stability. Within the linearized PB formalism, both domains appear equally optimized: the mutation free energy ($\Delta G^\circ_{\mathrm{mut}}$) of simultaneously replacing all protein sidechains with hydrophobic isosteres is $-1.4 \pm 4.6\mathrm{kcal\,mol^{-1}}$ and $+0.5 \pm 5.9$ for MVL_N and MVL_C.

Insights into the oligomeric nature of $\mathbf{MVL_C}$ Analytical ultracentrifugation experiments revealed that MVL_C dimerizes at µM concentrations (Figure 3-5C). Based on the structure of

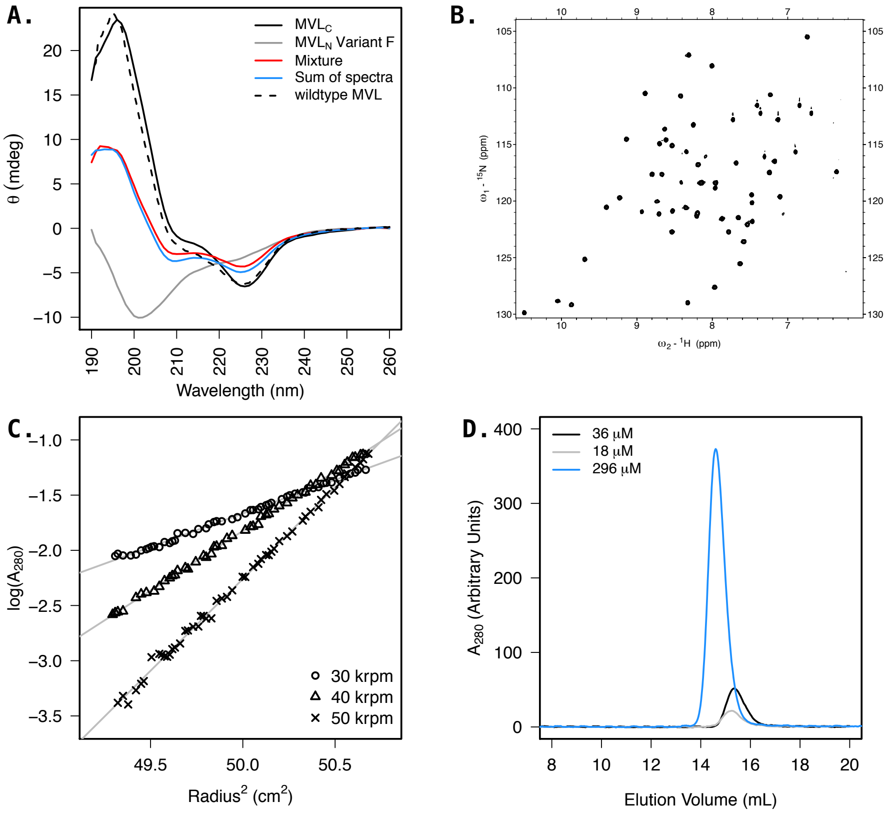


Figure 3-5: **A**. Far-ultraviolet CD spectra of MVL_C and MVL_N (black and grey lines) reveal that while the former possesses helical structure, the latter is unfolded. We tested whether hetero-association induces the folding of unstructured MVL_N; however, the spectrum of the mixture of the two variants (red) is identical to that of the weighted sum of their respective spectra (blue). **B.** The HSQC spectrum of MVL_C is well dispersed in both ^{1}H and ^{15}N dimensions, confirming that the protein is adopts a compact structure. The spectrum shown here was collected at 0.2 mM protein concentration; spectra collected upon two- and four-fold dilution are identical to the one shown. **C.** AUC sedimentation equilibrium experiments revealed that MVL_C behaves as an ideal dimer at 36 μM, with an apparent buoyant MW of 11 992 Da (expected MW of 12 059 Da). **D.** Analytical gel filtration chromatography shows that MVL_C elutes as a single monodisperse peak at all concentrations.

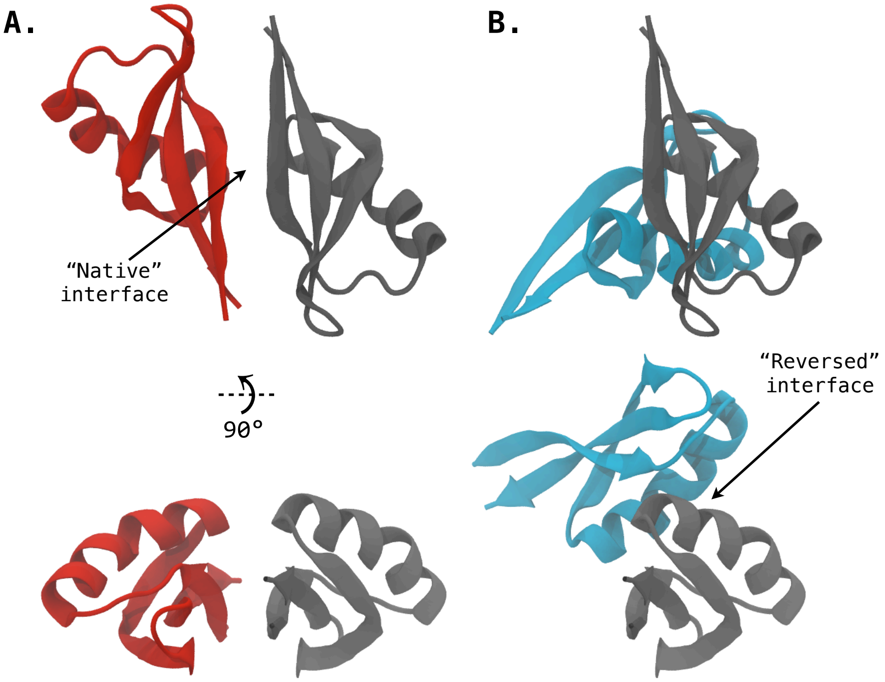


Figure 3-6: **A**. In wildtype MVL, the two C-terminal domains interact through a hydrophobic "native" interface involving the edges of two β-sheets. **B.** In wildtype MVL, the N- and C-terminal domains interact through a smaller interface, which involves electrostatic interactions between two α helices. Shown are two $\mathrm{MVL_C}$ domains, oriented to interact through the "reversed" interface. These two modes of binding correspond to schematic illustrations in Figure 3-3A.

wildtype MVL, there two possible dimerization modes of $\mathrm{MVL_C}$ (Figure 3-6). The first mode, which we refer to as "native" relies on a hydrophobic interface between the two molecules; this interface is used between the two C-terminal domains of full-length MVL. The second mode of binding, which we refer to as "reversed", buries a smaller amount of surface area, but is stabilized by salt bridge interactions (Figure 3-6 and Table 3.2). We constructed computational models of the two $\mathrm{MVL_C}$ dimers, and subjected them to MD simulation. Both complexes remained associated throughout the course of the simulation; MM/PBSA binding energy calculations revealed that the "native" interface is preferred energetically (Table 3.2). This mode of binding buries 360 Å^2 of additional surface area, compared to the "reversed" mode, and as a result makes more favorable vdW interactions. Surprisingly, the reversed binding mode features more favorable electrostatic interactions between the two polypeptide chains ($-2\,\mathrm{kcal\,mol^{-1}}$ *vs.* $10\,\mathrm{kcal\,mol^{-1}}$ for $\mathrm{MVL_N}$ and $\mathrm{MVL_C}$, respectively). We decomposed the electrostatic binding free energy $\Delta G^{\circ}_{\mathrm{elec}}$ into the desolvation penalty paid by the two chains and the direct (intermolecular) and indirect (intramolecular) electrostatic interaction free energies. The bottom of Table 3.2 shows the results: the reversed binding mode enables direct electrostatic interactions of $-39.4\,\mathrm{kcal\,mol^{-1}}$ between the two chains; however, each chain pays a desolvation penalty of approximately $20\,\mathrm{kcal\,mol^{-1}}$.

In order to experimentally determine which binding mode is engaged in solution, we tested mutants of MVL$_C$. The variant T64D was designed to disrupt the native interface, and the variant L80K was designed to disrupt the reversed interface. However, the yield of both proteins was drastically reduced upon mutation, and neither could be purified to levels required for biophysical characterization.

Binding Mode	ΔSASA	$\Delta G^\circ_{\mathrm{vdW}}$	$\Delta G^\circ_{\mathrm{h}\phi}$	$\Delta G^\circ_{\mathrm{elec}}$	$\Delta G^\circ_{\mathrm{bind}}$
Native	-1536.4	-59.2 ± 6.7	-6.8 ± 0.5	$+9.9 \pm 2.5$	-56.1 ± 8.4
Reversed	-1206.5	-38.9 ± 4.3	-5.2 ± 0.2	-2.0 ± 4.4	-46.0 ± 6.2

Binding Mode	Desolvation	Indirect	Direct	$\Delta G^\circ_{\mathrm{elec}}$
Native	$+11.1 \pm 1.7$	-1.4 ± 0.8	-9.4 ± 3.0	$+9.9 \pm 2.5$
	$+11.3 \pm 1.5$	-1.8 ± 0.9	-9.4 ± 3.0	$+9.9 \pm 2.5$
Reversed	$+20.5 \pm 3.7$	-1.2 ± 0.8	-39.4 ± 7.0	-2.0 ± 4.4
	$+19.6 \pm 2.6$	-1.5 ± 0.8	-39.4 ± 7.0	-2.0 ± 4.4

Table 3.2: **MVL$_C$ dimerization binding free energies calculated with the MM/PBSA approach**. The Native and Reversed binding modes correspond to MVL$_C$ dimers using one of the two available binding interfaces; in wildtype MVL, MVL$_C$:MVL$_C$ interactions are made through the mostly-nonpolar beta-strand interface, and interactions between MVL$_C$ and MVL$_N$ are made through a helix–helix interface (Figure 3-6). *Top.* The MM/PBSA-derived binding energetics of MVL$_C$ dimers in two binding modes. All energy values are given in kcal mol^{-1}. The Native binding mode is nonpolar in nature, buries 330 Å^2 of additional solvent-accessible surface area, and makes more favorable van der Waals interactions. The Reversed binding mode makes more favorable electrostatic interactions by virtue of intermolecular salt bridge interactions. *Bottom.* Decomposition of Poisson–Boltzmann ($\Delta G^\circ_{\mathrm{elec}}$) energetic contribution to binding into the desolvation penalty paid by, and the intra- (Indirect) and inter-molecular (Direct) interactions made by each chain. The Reversed binding mode results in more favorable direct interactions, and higher desolvation penalties, compared to the Native binding mode; values are given for both chains as a measure of precision of our approach.

3.4.3 MVL$_C$ Neutralizes HIV with Potency Equal to the Wildtype Protein

Having confirmed that the C-terminal domain adopts a compact fold similar to the wildtype protein, we investigated whether the isolated domain is able to inhibit infection by HIV. Four Env-pseudotyped viruses from laboratory-adapted subtype B HIV strains were treated using wildtype MVL and MVL$_C$ as potential inhibitors (Figure 3-7). MVL$_C$ neutralized HIV with potency equal to that of the wildtype protein for all four viral strains. In order to compare the tetra- and monovalent inhibitors directly, we present the concentration of each inhibitor as the number of binding sites (equivalent to normalizing by MW). While in some of the assays (JRCSF, HxB2, 89.6, shown in Figure 3-7) it may appear that MVL$_C$ is a more potent inhibitor, at the present time we cannot conclude this with absolute certainty. The inherent noise of the cell-based inhibition assays, in addition to the fact that 100% infectivity in not reached in three out of the four assay with low (pM) inhibitor concentrations, lead us to conclude that the two inhibitors are indistinguishable in their potency.

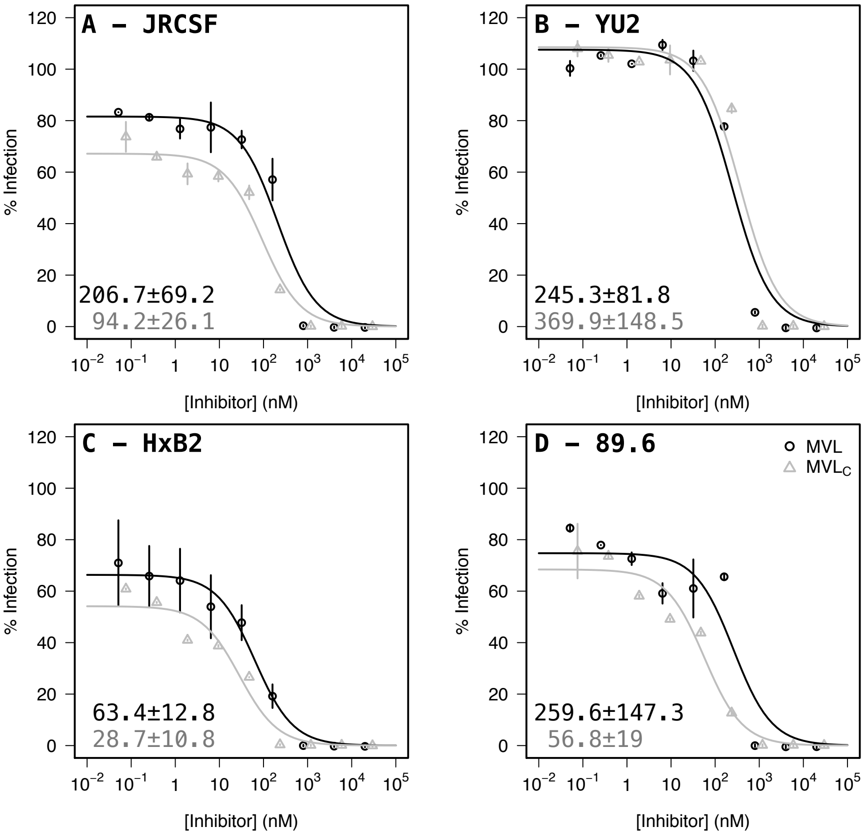


Figure 3-7: **Antiviral activity of MVL and MVL$_C$ in an Env-pseudotyped HIV neutralization assay**. Dose-response curves for antiviral activity of MVL$_C$ and full-length wildtype MVL. Both proteins were tested as inhibitors of infection by HIV pseudotyped with Env glycoprotein from four laboratory-adapted subtype B HIV strains (**A**: JRCSF, **B**: YU2, **C**: HxB2, **D**: 89.6). Inhibition of MVL is shown in black circles, and of MVL$_C$ in grey triangles. Vertical lines represent standard deviation error bars calculated from three experiments. The IC$_{50}$ values of each inhibitor against a particular HIV strain are provided in each panel. Protein concentrations were normalized to the concentration of binding sites (i.e. MVL dimer concentrations were divided by four).

3.5 Discussion & Conclusions

In this work, we established that MVL is not an ideal starting point for engineering a combinatorial scaffold for screening carbohydrate-binding specificity. The N-terminal domain of MVL was unfolded under all experimental conditions, and all efforts to stabilize the domain failed. We believe that a number of factors determine the stability difference between the two domains: first, two of the four positive-ϕ angle positions of MVL$_N$ are not glycine; second, the N-terminal domain is more loosely packed than $\mathrm{MVL_C}$; third, the dimerization interactions of $\mathrm{MVL_N}$ may be of lower affinity than those of $\mathrm{MVL_C}$, and as a result shift the equilibrium. While we attempted a number of approaches to induce folding of $\mathrm{MVL_N}$, we were not successful. Several additional strategies were not attempted: these include introducing mutations to facilitate disulfide-bond formation between positions 22/48 or between 26/50 (Figure 3-1C), or improving the packing of $\mathrm{MVL_N}$ using rational or computational approaches. For example, position 48 (valine) appears able to accommodate an additional methyl group; many domains of the MVL family feature an isoleucine at this position.

The asymmetry between the stabilities of the two domains of MVL is surprising. Since both domains are folded when part of the wildtype protein, we anticipate that the observed *in vitro* instability of $\mathrm{MVL_N}$ is overcome through enhanced local concentration of the two domains after the wildtype protein is synthesized. For instance, if the C-terminal domains fold spontaneously when part of wild-type MVL, while the N-terminal domains do not, the association of the C-terminal domains may bring the unstructured N-terminal domains together and induce their association and folding. Attempts to do this by mixing individual domains failed, likely because the domains were not tethered by the five-residue linked which joins them in the wildtype molecule. Kinetic stopped-flow folding experiments may reveal whether this type of configurational cooperativity is responsible for inducing the folding of the amino-terminal domain.

To our knowledge, our study presents the first example of reducing the valency of a multivalent HIV inhibitor which leads to no loss of neutralization potency. It has been previously shown that well-characterized HIV-inhibiting lectins CVN or griffithsin lose antiviral potency upon monomerization [122,129]. On unresolved question regarding the mechanism of MVL function is the identity of the *N*-linked carbohydrates targeted by the lectin. MVL recognizes the GlcNAc_rich core common to both high-mannose and complex oligosaccharides. Thus, the lectin may bind to multiple types of gp120 glycans. An intriguing extension of this work is the design of mutants which would shift the $\mathrm{MVL_C}$ dimer equilibrium by destabilizing the dimerization interface. We evaluated several mutants with this goal in mind; however, all were were unstable, possibly because of the drastically different sidechain properties upon mutation. Another possible approach which would further address issues of MVL multivalency is mutagenesis of the wild type protein to knock out carbohydrate binding in the amino- or carboxy-terminal domains. During the course of computational studies of carbohydrate recognition by MVL (Chapter 4) we uncovered an unprecedented geometric asymmetry between carbohydrate recognition by the two domains. We may thus expect that the two knockouts (simultaneously abolishing binding by the amino- or carboxy-terminal domains) may differ in their potency because of differences in their carbohydrate binding geometries.

Chapter 4

Structural Modeling & Molecular Dynamics Studies Provide Insight into Carbohydrate Recognition by MVL[*]

Abstract

Microcystis viridis Lectin (MVL, UniProt ID: Q9RHG4) is a 24-kDa homodimeric protein which inhibits infection by the human immunodeficiency virus (HIV) by binding to the viral glycoprotein gp120. Unlike other well-characterized HIV-inhibiting lectins such as cyanovirin-N or griffithsin, MVL is specific for the GlcNAc-containing core of *N*-linked oligosaccharides. Understanding the energetics of coordinating these groups will provide a more complete understanding of lectin–carbohydrate interactions in the context of HIV.

In this work, we present detailed analysis of carbohydrate recognition by MVL. Relying on molecular modeling, extensive molecular dynamics simulations, binding free energy calculations and energetic decomposition we elucidate the source of high-affinity binding by MVL to a diverse sugar panel featuring high-mannose and chitin-derived oligosaccharides. We uncover several novel features of MVL carbohydrate recognition, including an asymmetry in the interactions made by the amino- and carboxy-terminal domains of MVL. We also describe a dual binding mode of the tetrasaccharide GlcNAc_4 to MVL, and re-analyze experimental data collected for this ligand. By providing an atom-level description of the binding event, our results complement previous experimental investigations of MVL, and underscore the need to consider the dynamics in protein–carbohydrate investigations.

4.1 Introduction

Structural studies of protein–carbohydrate complexes elucidate the bound state at atomic resolution. However, the two most common techniques used in structural elucidation, X-ray crystallography and NMR restraint-based docking, suffer from limitations. Crystallography provides a static snapshot of the average conformation, and is often performed at low temperature, thereby dampening molecular fluctuations. NMR, being a solution technique, can be applied under physiologically-

[*] The continuum van der Waals calculations presented in this chapter were performed by Dr. Yukiji K. Fujimoto.

relevant conditions and yields information about the bound-state ensemble. Sometimes, however, the restraints generated through the use of Nuclear Overhauser Effect (NOE) information are inadequate to unambiguously position a ligand in the ideal geometry [58] (and work described herein). Intermolecular restraints used to orient the ligand also may lack peripheral interactions with high signal-to-noise ratio or high dynamic content. Additionally, NMR-based refinement often makes use of a crude potential which fails to capture the fine details of the bound-state energetics [156].

The binding affinity of a receptor for its cognate ligand is a property of a dynamic configurational ensemble, and for most ligand/receptor complexes it is impossible to calculate accurate binding energies using a single conformation. Neither experimental structural technique discussed above generates an extensive bound-state ensemble. However, both structural techniques can provide a starting point for molecular dynamics (MD) simulation, a computational technique complementary to experiment [44, 57, 58]. Starting from an initial configuration of atoms, MD simulation engines generate conformational ensembles which are typically more dynamic than seen in crystallographic studies. These ensembles can be interrogated to yield the free energy of binding [96, 172], the per-residue contributions to binding [57], structural details of the binding event, as well as countless time-dependent structural properties or correlation functions. First-principles molecular dynamics approaches can also be used to refine experimental data or low-resolution models [149].

In order to understand lectin function in its native or non-native context, questions related to specificity, binding affinity, and the role of multivalency need to be addressed; a number of mature technologies exist to address some of these questions. Glycan microarrays can address lectin binding specificity by screening hundreds of diverse natural and synthetic carbohydrates [1, 22, 53]. NMR-monitored titration [166], in addition to other structural techniques, provides insight into which portion of the receptor is involved in binding a certain polysaccharide fragment. Binding affinity for carbohydrates can be measured by isothermal titration calorimetry (ITC) or surface plasmon resonance. Under the right set of experimental conditions, enzyme-linked or fluorescence-based assays can also be used.

Computational investigations allow an atom-level view of the protein–carbohydrate recognition event. Upon binding their cognate receptors, carbohydrates typically bury large surface areas, are in contact with multiple protein residues, and are coordinated by numerous hydrogen bonds within the binding site. However, the contributions of molecular groups to binding are unequal in magnitude due to variation of their functional group, distance, or dynamic content within the bound-state ensemble [57]. Decomposition of binding affinity into per-residue contributions is impossible to address experimentally, but is trivial within the pairwise-decomposable framework of the molecular mechanics potential [23, 24] or the linearized Poissson–Boltzmann formalism [76]. Additionally, MD simulations enable the investigation of ligands which may be challenging or costly to obtain in quantities required for structural or biophysical characterization, and facilitate the design of mutations with the goal of improved binding affinity or altered specificity. Ligands of weak affinity, which are beyond experimental detection, can be routinely studied by MD. This is because the dissociation kinetics of weak-binding ligands are often slower than the timescales explored during the simulation time course.

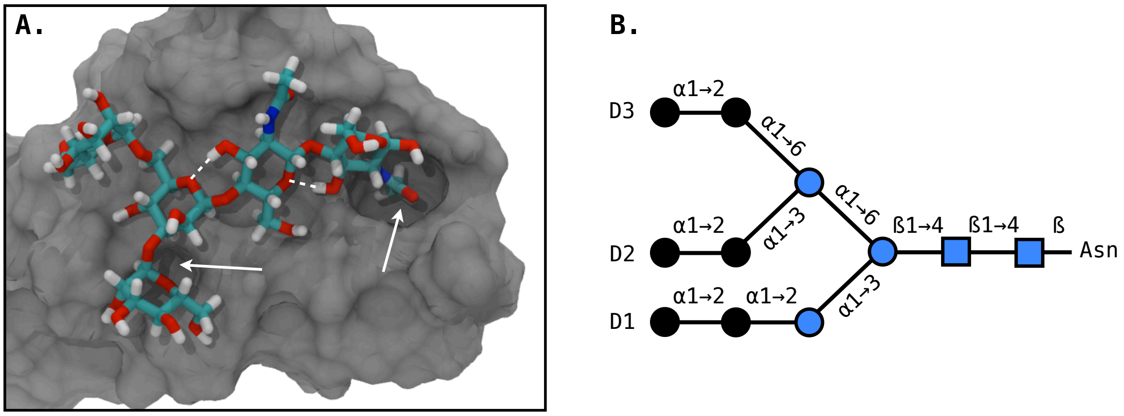


Figure 4-1: **Carbohydrate recognition by MVL. A**. A single MVL domain in complex with Man$_3$GlcNAc$_2$, the Y-shaped core substructure of high-mannose and complex *N*-linked oligosaccharides. The two deep cavities on the MVL surface are indicated with white arrows. The *pseudo*-rings formed between adjacent $\beta(1 \to 4)$-linked sugar residues are indicated with dashed lines. Figure rendered with VMD [84] based on PDB entry `1ZHS` [188]. **B**. Schematic representation of Man$_9$GlcNAc$_2$ undecasaccharide (also known as Man$_9$), linked to a glycoprotein through an Asn sidechain. Man residues are represented by ●, and GlcNAc by ■. The Man$_3$ pentasaccharide displayed in **A** is shaded blue.

In this work, we investigated carbohydrate binding by the HIV-inhibiting protein MVL, a lectin which was originally isolated from the freshwater blue-green algæ *Microcystis viridis* [196]. Previous structural and biophysical studies revealed that the protein is an obligate homodimer which contains four carbohydrate binding domains [16, 188]. The exact mechanism of viral inhibition by MVL is unknown, but it is thought to involve interactions with the core substructure of oligosaccharides which decorate the surface of the viral glycoprotein gp120 [16]. These carbohydrates, termed the “glycan shield”, obscure the underlying viral protein epitopes from detection by the humoral immune system. Antiviral strategies which target the glycan shield interfere with adhesion, the initial step of the viral life cycle [151].

The use of glycosylation by HIV is atypical when compared to the effect this post-translational modification has on human proteins. One distinguishing feature of viral glycoproteins from those of human origin is the abundance of Man$_9$GlcNAc$_2$ and other high-mannose glycan moieties on the viral surface; this glycan is rare on human glycoproteins, since it is normally heavily remodeled in the Golgi apparatus to yield shorter substructures, some of which are extended to complex glycans. It is not clear why the HIV-1 glycans are not processed. One hypothesis is that the rate of viral glycoprotein synthesis, in addition to the high number of N–X–S/T glycosylation motifs may be responsible. However, it remains possible that the virus has either evolved a way to accelerate the export of its proteins before processing can happen, or a way to inhibit processing enzymes such as $\alpha(1 \to 2)$-mannosidases found in the endoplasmic reticulum.

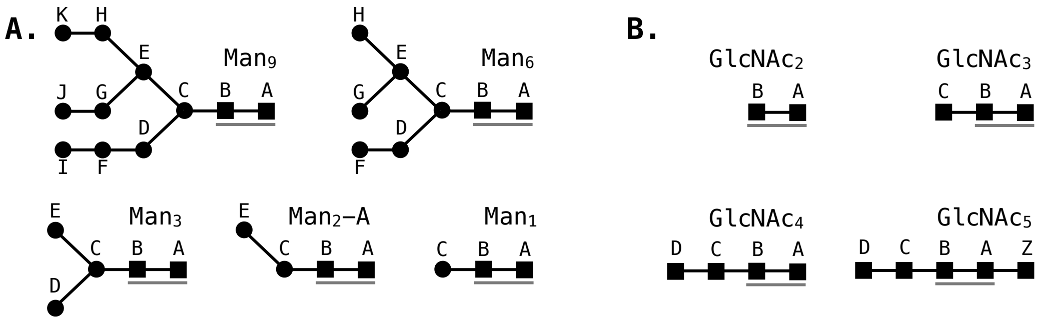


Figure 4-2: **Carbohydrate ligands discussed in the present work.** Mannose (●) and N-acetyl-glucosamine by (■) residues are identified with capital letters for each carbohydrate model. The reducing-end residue (ring A, unless Z is present) is oriented to the right. The GlcNAc$_2$ motif (rings A and B) common to all ligands and necessary for binding MVL is underlined. **A**. The branched high-mannose undecasaccharide Man$_9$ is shown; all high-mannose ligands investigated in this study are substructures of Man$_9$. **B**. A series of linear chitin-derived oligosaccharides, extended from the core GlcNAc$_2$ motif.

A number of plant or algal lectins inhibit viral infection, typically by binding to mannose-rich portions of Man$_9$. The well-characterized antiviral lectins cyanovirin-N [15] and griffithsin [2, 200], as well as the neutralizing antibody 2G12 [182], recognize the terminal arms of Man$_9$GlcNAc$_2$. MVL, on the other hand, recognizes the core of N-linked carbohydrates, and its X-ray structure was determined in complex with Man$_3$GlcNAc$_2$ (Figure 4-1) [16]. While the binding of lectins to mannose residues in the context of HIV is fairly well understood, recognition of their GlcNAc core involves a distinct set of interactions and molecular geometries, and understanding the energetics, and well as the geometries of these interactions may aid in understanding the mechanism by which lectins inhibit viral infection, or in design of improved inhibitors.

4.2 Materials & Methods

Construction of Starting Protein:Carbohydrate Models. The coordinates of full-length MVL were taken from PDB entry 1ZHS. The crystal asymmetric unit contains four MVL homodimers in complex with sixteen molecules of Man$_3$GlcNAc$_2$. Coordinates for protein chains A and B, as well as the bound carbohydrate ligands, were retained. Protonation states of histidine residues, as well as the flip states of asparagine or glutamine, were determined with REDUCE [194]. Truncated carbohydrate models were constructed by deleting the appropriate sugar residues; additional residues in models extended from Man$_3$ were constructed from internal coordinates using CHARMM [27]. In the cases where a clash was observed, the clashing groups were energy minimized while the remaining groups were fixed in their original orientation. For all carbohydrate models, the hydroxyl group of the reducing-end anomeric carbon was in the β configuration.

Molecular Dynamics Simulation. All system setup and post-processing steps were performed with CHARMM [27]. Individual domains in complex with carbohydrates were solvated in an orthorhombic box containing approximately 3300–4300 TIP3P water molecules (≈9200 for MVL:Man$_9$ simulation), and charge-neutralized by adding K^+ and Cl^- ions to a final concentration of 150 mM. To avoid self-interactions in the periodic cell, water box dimensions were set such that the solute was 10 Å away from the cell side. Solvation and neutralization steps used a locally-modified input script originally retrieved from www.charmmtutorial.org. In rare cases when visual inspection identified water molecules inside the protein, the offending molecules were translated into the bulk solvent. The CHARMM22-CMAP [115] and CSFF [97] molecular mechanics force fields were used for the protein and carbohydrate components, respectively.

Isothermal–isobaric (NPT) ensemble molecular dynamic simulations were carried out at 300 K with the NAMD 2.6 engine compiled for the IBM Blue Gene/L architecture [143]. Constant pressure of 1 atm was maintained with the Nosé–Hoover Langevin piston pressure control. All bonds involving hydrogen atoms were constrained to their equilibrium lengths using the SHAKE algorithm. The simulation time step was set to 2 fs and non-bonded interactions were evaluated every step. We used a cutoff of 12 Å for all Lennard–Jones and short-range electrostatic interactions. Long-range electrostatics were treated using a Particle Mesh Ewald method with a fourth-order interpolation scheme. The solvated and neutralized system was minimized for 5000 steps and heated to the production temperature over the course of 200 ps. No restraints were employed during the equilibration or production steps.

Binding Free Energy Calculation. Binding free energies were computed with the MM/PBSA model [96, 172]. The total binding energy for each snapshot was computed as the sum of a Poisson–Boltzmann-based electrostatic contribution ($\Delta G^\circ_{\text{elec}}$), the intermolecular van der Waals energy ($\Delta G^\circ_{\text{vdW}}$), and a term proportional to the solvent accessible surface area buried on binding. In the latter case, the area was computed with CHARMM using a 1.4 Å radius probe and the energetic contribution was given by $\Delta G^\circ_{\text{h}\phi} = 0.005 \cdot \Delta A + 0.86\,\text{kcal mol}^{-1}$ for each snapshot. Binding energies were calculated using snapshots extracted from the solvent-stripped MD trajectories at an interval of 1 ns.

PB Calculations. The linearized Poisson–Boltzmann equation was solved using a multi-grid finite-difference solver distributed with the Integrated Continuum Electrostatics (ICE) software suite (courtesy of Prof. B. Tidor), using standard protocols [28, 67, 77]. Atomic partial charges were from the CHARMM22-CMAP [115] and CSFF [97] parameter sets. The dielectric boundary was defined by the molecular surface generated with a 1.4 Å radius probe, and a 2.0 Å ion exclusion layer was used; the surfaces were generated using radii optimized for use in continuum electrostatic calculations [65, 135]. The solute and solvent dielectric constants were set to 2 and 80, respectively, at an ionic strength of 145 mM. Boundary conditions were computed using a three-step focusing procedure on a 129^3-unit cubic grid, with the molecule first occupying 23%, then 92%, and finally 184% of the grid. Boundary conditions at each focusing level were taken from the previous calculation, with Debye–Hückel potentials used at the boundary of the lowest focusing level. The highest-resolution grid was centered on the oligosaccharide, and potentials at atoms falling off this grid were taken from the middle-resolution calculation. Electrostatic contributions to the binding free energies were computed as the sum of a desolvation penalty for both the protein and the sugar and a bound-state, solvent-screened interaction.

Component Analysis. Electrostatic component analysis was performed with the ICE software package using standard protocols [28, 67]. Each protein residue was partitioned into three groups: carbonyl, amino, and side chain. Sugars were partitioned into one group per hydroxyl; in the case of GlcNAc, group definitions included the *N*-acetyl moiety. For each group, the desolvation penalty, inter- and intramolecular interactions were computed. We define the sum of these as the mutation free energy $\Delta\Delta G^\circ_{\text{mut}}$, equivalent to the difference in free energy between the natural system and a hypothetical mutant with that group replaced by a hydrophobic isostere (in the context of all other groups in their natural state) [77]. Equivalent group definitions were also used to calculate the van der Waals interactions between carbohydrate and protein groups, and the contribution of each group to binding. The CHARMM potential was employed for van der Waals energy evaluation [116].

4.3 Results

Using explicit-solvent molecular dynamics simulations, we investigated the binding of a broad panel of carbohydrate ligands to MVL. The complete list of ligands in this study, the nomenclature employed herein, as well as the simulation length for each protein–carbohydrate system investigated herein are given in Table 4.1.

Abbreviation	Composition	Receptor(s)	MD Run Length (ns)
Man_1	$Man_1GlcNAc_2$	MVL	400
Man_2-A	$Man_2GlcNAc_2$	MVL domains	2 × 500
Man_3	$Man_3GlcNAc_2$	MVL domains, MVL_C-E101R	2 × 400, 200
Man_3-Asn	$Man_3GlcNAc_2Asn$	MVL_C	300
Man_6	$Man_6GlcNAc_2$	MVL domains	2 × 500
Man_9	$Man_9GlcNAc_2$	MVL domains, MVL	2 × 700, 200
$GlcNAc_2$	$GlcNAc_2$	MVL_C	400
$GlcNAc_3$*	$GlcNAc_3$	MVL_C	2 × 400
$GlcNAc_4$*	$GlcNAc_4$	MVL_C	2 × 500
$GlcNAc_5$	$GlcNAc_5$	MVL_C	650

Table 4.1: **Summary of molecular dynamics simulations performed as part of this study.** Seventeen independent trajectories were generated, with cumulative simulation time of 8.15 µs. Structures of all the carbohydrate ligands are given in Figure 4-2.
*$GlcNAc_3$ and $GlcNAc_4$ were simulated in two alternative binding modes. The details of this are discussed in the text.

4.3.1 High-mannose Saccharide Binding

Initial MD studies of the tetravalent MVL:Man_3 complex proved unstable, as one of the four ligands dissociated after only 50 ns of simulation; this simulation was terminated. The failure of this initial approach is likely due to artifactual or non-ideal system setup, since Man_3 is a known experimental binder of MVL, and remained bound in subsequent simulations (*vide infra*).

Man_1. Since initial MD studies with the Man_3 ligand ran into difficulties, we truncated the carbohydrate model, removing the two terminal Man residues to yield the trisaccharide Man_1. While this ligand has not been experimentally determined to bind MVL, in simulation it remained bound in all four sites, and occupied two closely-related conformations. The reducing-end GlcNAc A remained anchored in the *N*-acetyl binding cavity, while GlcNAc B was closely associated with the protein. Man C populated two alternative conformations. Clustering of the combined B–C glycosidic linkage data set for the four ligands using the *k*-means algorithm ($k = 2$) revealed clusters of staggered and extended conformations, which contained 87% and 13% of the population, respectively. In the more abundant staggered conformation, the glycosidic ϕ (O5–C1–O–C$(x)'$) and ψ (C1–O–C$(x)'$–C$(x\text{–}1)'$) angles between rings B and C are $-158.2 \pm 11.7°$ and $91.3 \pm 6.7°$, respectively; the smaller cluster adopted extended conformations with ϕ and ψ of $-77.6 \pm 11.9°$ and $117.1 \pm 13.5°$. The A–B glycosidic linkage populated an extended conformation similar to that of the smaller cluster of the B–C linkage (ϕ and ψ of $-75.5 \pm 6.4°$ and $109.7 \pm 6.9°$).

The two bound-state conformations of Man_1 are distinguished by the presence of a hydrogen bond between the Man C2 hydroxyl and the carbonyl of Trp37/Trp96 (residue numbering in the N- and C-terminal domains, respectively). In mannose residues the C2 hydroxyl is equatorial, and the sugar residue must adopt a staggered conformation in the binding site of MVL in order to make this contact. The interaction was engaged in the staggered conformation, and disengaged in the extended conformation; in the C-terminal domain of MVL, the average distance between carbonyl oxygen of Trp96 and the C2 hydroxyl of Man was 1.9 ± 0.3Å and 5.6 ± 1.1Å for the two conformations, respectively.

We calculated the binding free energy for Man_1 interacting with each binding site of MVL. In all domains, the ligands bury a comparable amount of surface area, and make comparable van der Waals interactions (Table 4.2). The majority of the binding energy ($-33\,\mathrm{kcal\,mol^{-1}}$) was derived from the molecular mechanics portion ($\Delta G^{\circ}_{\mathrm{vdW}} + \Delta G^{\circ}_{\mathrm{h}\phi}$) of the MM/PBSA equation, with electrostatics contributing weakly. PB electrostatic calculations suggested a small difference between the two domains of about $1\,\mathrm{kcal\,mol^{-1}}$, with the $\mathrm{MVL_C}$ domains interacting more favorably with the ligand.

Role of intra-ligand interactions in $\beta(1\rightarrow4)$-linked polysaccharides. The extended conformation is frequently encountered in $\beta(1\rightarrow4)$-linked polysaccharides, and was commonly seen in our simulations. In this conformation, glycosidic ϕ/ψ torsional angles of $-80°/110°$ orient the polysaccharide in a pleated-like sheet. The distinguishing feature of this conformation is that the

C3 hydroxyl of one residue donates a hydrogen bond to the ring O5 atom of the preceding residue, creating a hydrogen-bond-mediated *pseudo*-ring between the two (Figure 4-1A). The extended conformation is common for $\beta(1 \rightarrow 4)$-linked polysaccharide polymers, such as chitin and cellulose. In chitin-derived GlcNAc polysaccharides, the extended conformation orients N-acetyl groups of adjacent residues in diametrically opposite directions [195].

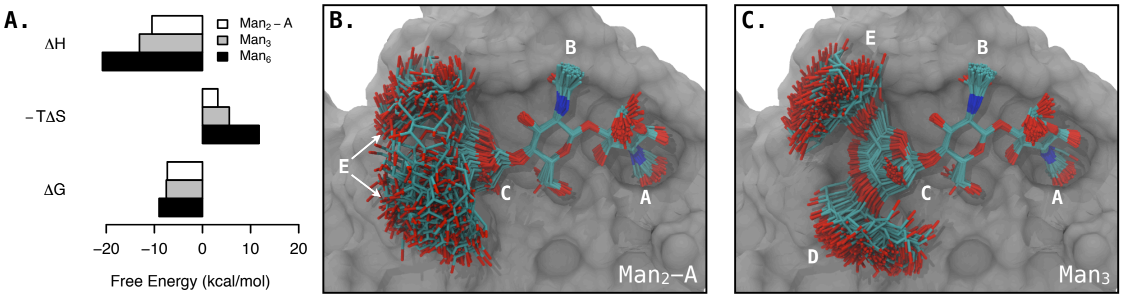


Figure 4-3: **Importance of dynamics in protein–carbohydrate binding. A**. Experimentally measured free energies, enthalpies and entropies of carbohydrate binding to MVL [16] ($\Delta G = \Delta H - T\Delta S$); the affinities of MVL for Man$_2$-A and for Man$_3$ are equal. The binding of larger saccharides is increasingly favorable enthalpically, but is opposed by increasing entropic penalties. Notably, compared to Man$_2$-A, the binding of the larger Man$_3$ involves enthalpic contributions of greater magnitude, but is opposed by a larger entropic component; these trends are further increased with Man$_6$. **B**. The bound-state structural ensemble of Man$_2$-A explores two main conformations (extended and staggered) which interconvert through a B–C glycosidic linkage flip. **C**. The structural ensemble of Man$_3$ is less diverse than that of Man$_2$-A, as ring D prevents the ligand from rotating about the B–C linkage while bound to the protein.

Man$_2$-A. For computational efficiency, we simulated all subsequent glycan models with the individual domains of MVL (exceptions noded). Binding of the tetrasaccharide Man$_2$-A to MVL has been previously characterized by ITC, and its affinity is equal to that of the larger Man$_3$ (Figure 4-3A) [16]. Simulations of Man$_2$-A revealed that the ligand is flexible about the B–C glycosidic linkage, adopting two conformations, similar to what was observed for Man$_1$ (Figure 4-3B). The populations of the staggered and extended conformations were unequal when comparing the two domains of MVL. In complex with MVL$_\mathrm{N}$ the ligand adopted the staggered and extended conformations 54 and 46% of the time, respectively. In complex with MVL$_\mathrm{C}$, the populations of the staggered and extended conformations were 29 and 71%.

The calculated binding energy for Man$_2$-A was -36.4 ± 3.5 kcal mol^{-1} and -35.9 ± 3.7 for MVL$_\mathrm{N}$ and MVL$_\mathrm{C}$, respectively (Table 4.2). The majority of the binding free energy (-35.5 kcal mol^{-1}) was derived from the nonpolar portion of the equation, with electrostatics contributing weakly. Compared to MVL$_\mathrm{N}$, MVL$_\mathrm{C}$ interacted more favorably electrostatically with the ligand, as was the case with Man$_1$. Man$_2$-A is the only ligand in this study which we observed to interact more favorably with MVL$_\mathrm{N}$ than with MVL$_\mathrm{C}$. In the case of binding Man$_2$-A, the difference in van der Waals interactions between the two domains is the consequence of the shift in the extended and staggered populations of the ligand.

$\mathbf{Man_3}$. We repeated the simulation of Man_3 in complex with individual MVL domains. In both simulations, the carbohydrate remained bound; its average conformation was similar to that in the MVL crystallographic ensemble [188]. As expected, the MD-derived ensemble was more diverse than that of the asymmetric unit (Figure 4-3), and each of the nine glycosidic torsional angles sampled wider regions of the ϕ/ψ dihedral space. The average ϕ/ψ for the A–B linkage were: $-71.6 \pm 6.8°/110.6 \pm 6.6°$; for the B–C linkage: $-71.5 \pm 7.0°/123.9 \pm 8.6°$; for the C–D linkage: $67.5 \pm 11.4°/174.7 \pm 16.9°$; for the C–E linkage: $76.3 \pm 16.7°/-112.6 \pm 21.7°$. The ω dihedral angle of the C–E linkage averaged $54.4 \pm 8.4°$. Man_3 was significantly less dynamic than Man_2-A in simulation (Figure 4-3C), as the presence of Man D prevents the B–C linkage flip observed for Man_1 and Man_2-A.

Man_3 buries approximately 200 Å^2 and 130 Å^2 of additional surface area upon binding MVL_C, compared to Man_1 and Man_2-A, respectively (Table 4.2). The calculated binding energy of Man_3 was -40.5 ± 3.3 and -41.6 ± 4.0 kcal mol^{-1} for MVL_N and MVL_C. The majority of the binding free energy was derived from the nonpolar portion of the equation, with electrostatic interactions opposing binding in both domains.

We additionally investigated interactions of Man_3 with Glu101Arg (E101R) mutant of MVL_C. This mutant was designed by Mr. Yiwei Cao using a computational protein design approach with the goal of optimizing Man_3 binding. We observed that the sidechain of Arg101 interacted favorably with Man_3, and that the carbohydrate buried 37 Å^2 of additional surface area upon binding, compared to wild type MVL_C. As predicted by protein design calculations, Man_3 bound E101R slightly mote favorably than wild type MVL_C, at -43.4 ± 5.0 kcal mol^{-1}; the improved binding free energy primarily arose from enhanced van der Waals interactions (Table 4.2).

$\mathbf{Man_3}$-Asn "Glycopeptide". While biophysical or thermodynamic studies of MVL investigate binding to isolated saccharides, the protein inhibits HIV by binding the glycoprotein gp120 [16]. Consequently, its interactions with N-linked oligosaccharides may be influenced by the presence of the Asn residue which links the saccharide to the protein or by the protein itself. We set out to investigate this, extending Man_3 through a β linkage by an asparagine residue; in our model, the N- and C-termini of the Asn were charged. While the presence of charged termini does not make this an optimal model of MVL binding to a glycoprotein, as part of our experimental characterization of MVL (Chapter 3) we aimed to experimentally characterize binding to this ligand, which can be extracted in milligram quantities from soybean flour [182].

Man_3-Asn buries approximately 100 Å^2 of additional surface area upon binding MVL_C, compared to Man_3 (Table 4.2), with calculated binding energy of -44.2 ± 4.0 kcal mol^{-1}. The majority of the binding free energy (-46.2 kcal mol^{-1}) was derived from the nonpolar portion of the equation, with unfavorable electrostatic contribution of 1.9 kcal mol^{-1}. The Asn residue of the glycopeptide

Ligand	Receptor	ΔSASA	$\Delta G^\circ_{\mathrm{vdW}}$	$\Delta G^\circ_{\mathrm{h}\phi}$	$\Delta G^\circ_{\mathrm{elec}}$	$\Delta G^\circ_{\mathrm{bind}}$
Man$_1$*	MVL$_\mathrm{N}$	−710.7	−30.4 ± 3.5	−2.7 ± 0.2	−1.6 ± 2.8	−34.7 ± 3.6
	MVL$_\mathrm{N}$	−714.1	−30.6 ± 2.8	−2.7 ± 0.1	−1.5 ± 2.6	−34.8 ± 3.0
	MVL$_\mathrm{C}$	−716.1	−30.2 ± 3.1	−2.7 ± 0.1	−2.9 ± 2.7	−35.9 ± 2.8
	MVL$_\mathrm{C}$	−719.2	−30.7 ± 3.0	−2.7 ± 0.1	−2.2 ± 2.8	−35.6 ± 3.0
Man$_2$-A	MVL$_\mathrm{N}$	−787.1	−33.4 ± 4.0	−3.1 ± 0.3	+0.2 ± 2.9	−36.4 ± 3.5
	MVL$_\mathrm{C}$	−767.4	−31.8 ± 4.0	−3.0 ± 0.3	−1.2 ± 3.2	−35.9 ± 3.7
Man$_3$	MVL$_\mathrm{N}$	−917.3	−38.6 ± 2.7	−3.7 ± 0.2	+1.8 ± 2.8	−40.5 ± 3.3
	MVL$_\mathrm{C}$	−925.2	−39.1 ± 3.4	−3.8 ± 0.3	+1.2 ± 2.8	−41.6 ± 4.0
	MVL$_\mathrm{C}$-E101R	−962.5	−40.3 ± 4.1	−4.0 ± 0.3	+0.8 ± 3.3	−43.4 ± 5.0
Man$_3$-Asn	MVL$_\mathrm{C}$	−1033.6	−41.9 ± 3.3	−4.3 ± 0.2	+1.9 ± 3.5	−44.2 ± 4.0
Man$_6$	MVL$_\mathrm{N}$	−1126.6	−46.2 ± 4.2	−4.8 ± 0.3	+3.6 ± 3.3	−47.4 ± 4.5
	MVL$_\mathrm{C}$	−1177.6	−49.0 ± 4.7	−5.0 ± 0.4	+2.3 ± 5.3	−51.7 ± 6.3
Man$_9$	MVL$_\mathrm{N}$	−1290.7	−50.6 ± 4.0	−5.6 ± 0.3	+5.7 ± 4.2	−50.5 ± 4.5
	MVL$_\mathrm{C}$	−1401.6	−56.0 ± 4.9	−6.1 ± 0.4	+5.3 ± 6.0	−56.8 ± 6.8
Man$_9$*,†	MVL$_\mathrm{N}$	−1297.2	−50.7 ± 4.2	−5.6 ± 0.3	+7.6 ± 4.5	−48.7 ± 4.3
	MVL$_\mathrm{N}$	−1294.4	−51.5 ± 4.2	−5.6 ± 0.3	+8.0 ± 4.2	−49.1 ± 4.0
	MVL$_\mathrm{C}$	−1410.3	−55.2 ± 4.6	−6.2 ± 0.3	+3.4 ± 6.1	−57.9 ± 5.7
	MVL$_\mathrm{C}$	−1364.6	−55.8 ± 4.5	−6.0 ± 0.4	+5.8 ± 5.1	−55.9 ± 5.7
GlcNAc$_2$	MVL$_\mathrm{C}$	−576.4	−24.7 ± 2.9	−2.0 ± 0.1	−1.3 ± 2.4	−28.0 ± 3.8
GlcNAc$_3$-A	MVL$_\mathrm{C}$	−774.1	−34.9 ± 3.2	−3.0 ± 0.2	+0.6 ± 2.7	−37.3 ± 3.6
GlcNAc$_3$-B	MVL$_\mathrm{C}$	−692.2	−27.3 ± 2.6	−2.6 ± 0.2	+0.5 ± 2.3	−29.4 ± 2.9
GlcNAc$_4$-A	MVL$_\mathrm{C}$	−868.6	−38.6 ± 6.6	−3.5 ± 0.4	+2.1 ± 2.5	−40.0 ± 7.4
GlcNAc$_4$-B	MVL$_\mathrm{C}$	−875.4	−36.8 ± 3.6	−3.5 ± 0.2	+1.9 ± 2.8	−38.4 ± 4.1
GlcNAc$_5$	MVL$_\mathrm{C}$	−929.8	−37.2 ± 7.1	−3.8 ± 0.5	+3.0 ± 2.7	−38.0 ± 8.0

Table 4.2: **Carbohydrate binding energetics calculated with the MM/PBSA approach**. Average change in solvent-accessible surface area lost upon binding (ΔSASA) is given in Å^2; ΔG° values are given in kcal mol^{-1}, with standard deviations provided as a measure of variation. Binding energetics were evaluated at 1 ns intervals. Calculating the binding free energy with increased frequency (every 0.1 ns) gave identical results.

*Binding values were extracted from simulations of wildtype MVL in complex with four ligands; average values are provided for each domain as a measure of precision. All other simulations were of individual MVL domains.

†For Man$_9$ in complex with tetravalent MVL, we observed carbohydrate–carbohydrate interactions between the N-terminal domain ligands; for this system MM/PBSA binding energetics were calculated using the appropriate protein–carbohydrate complex as the receptor.

interacted favorably with MVL, contributing -4.2 ± 1.9 kcal mol^{-1} to the van der Waals portion of the binding free energy.

Man$_6$. Currently, Man$_6$ is the tightest-binding monovalent MVL ligand known (Figure 4-3A); however, no structural data of how this ligand interacts with MVL is available [16, 188]. We extended Man$_3$ by three Man residues (rings F, G, H). Interestingly, when the $\alpha(1\rightarrow 2)$-linked Man F was appended to Man D, the former was placed within a small pocket on the surface of MVL$_N$ and MVL$_C$. This binding pocket (surrounded by Asp65, Gln95, and Arg97 in MVL$_C$, for instance) is observed in all carbohydrate complexes investigated in this study; however, Man$_6$ and Man$_9$ (discussed below) are the only ligands which occupy it.

The pentasaccharide core of Man$_6$ (rings A–E, equivalent to Man$_3$) was rigid during the course of the simulation, with measured ring heavy-atom RMSD of 0.5 Å; the nine glycosidic torsion angles common to Man$_3$ and Man$_6$ behaved identically. Man$_6$ buries approximately 250 Å^2 of additional surface area upon binding MVL$_C$, compared to Man$_3$ (Table 4.2). The calculated binding energy for Man$_6$ was -47.3 ± 4.5 kcal mol^{-1} and -51.7 ± 6.3 kcal mol^{-1}, for MVL$_N$ and MVL$_C$. The majority of the binding free energy (-51.0 and -54.0 kcal mol^{-1}) was derived from the molecular mechanics portion of the MM/PBSA equation, with electrostatics contributing unfavorably. We found that the enhancement of $\Delta G^{\circ}_{\text{vdW}}$ of Man$_6$ compared to Man$_3$ resulted from the interactions of mannoses F and H. In particular, ring F contributed -7.7 kcal mol^{-1} to binding (Table 4.3), by virtue of its tight intermolecular interactions with the Asp65/Arg97 binding pocket.

Man$_9$. Having confirmed that Man$_6$ behaves reasonably in simulation in complex with both MVL domains, we constructed models of the physiologically-relevant Man$_9$. This modification, along with its D2-truncated derivative Man$_8$, is abundant on gp120. We performed simulations of Man$_9$ in complex with individual domains of MVL, and with the tetravalent wildtype protein.

For the simulations of monomeric domains in complex with Man$_9$ we observed that MVL$_C$ interacts more tightly with Man$_9$ than does MVL$_N$, by virtue of more optimal van der Waals interactions and PB electrostatic complementarity (Table 4.2). The calculated binding energy was -50.5 ± 4.5 kcal mol^{-1} and -56.8 ± 6.8 kcal mol^{-1}, for MVL$_N$ and MVL$_C$, respectively. Unexpectedly, the model of tetravalent MVL in complex with Man$_9$ revealed carbohydrate–carbohydrate interactions between the D1 arms of Man$_9$ molecules occupying the N-terminal domains 4-4A). The magnitude of van der Waals sugar–sugar interactions is small ($\Delta G^{\circ}_{\text{vdw}}$ of -0.9 kcal mol^{-1}) in comparison with the interactions each ligand makes with the protein.

4.3.2 *N*-acetyl-chitooligosaccharide Binding

In addition to binding the GlcNAc core of branched *N*-linked glycans, MVL possesses modest glycosidase activity against linear chitin-derived poly-GlcNAc oligosaccharides [160]. The dual carbohydrate specificity within the binding site of MVL, coupled to the protein's enzymatic activity is highly unusual. However, since nothing is known about the role of MVL within its native

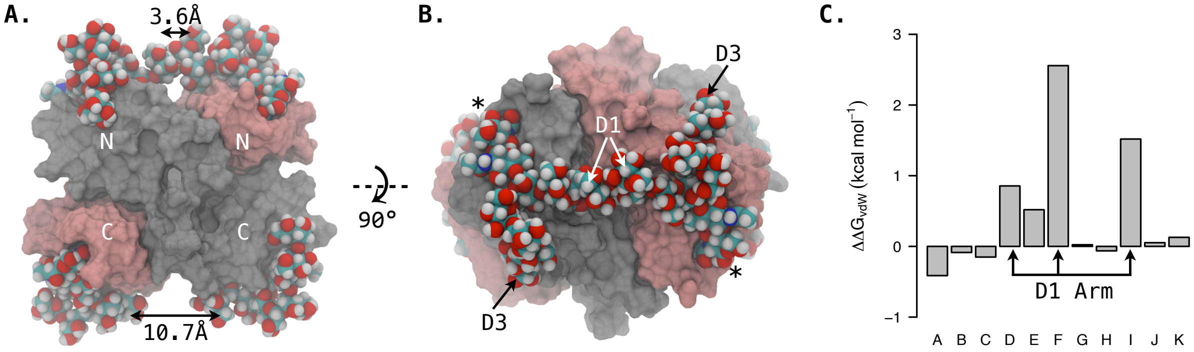


Figure 4-4: **Carbohydrate–carbohydrate interactions in the MVL:Man$_9$ complex. A.** The sugars bound in the two N-terminal domains of MVL come into physical contact during the simulation; the average minimum distance between the two is 3.6 Å. No such interactions are observed for the C-terminal domain ligands. **B.** Pairwise interactions between D1-arm Man residues F and I in the two ligands account for 99% of the intermolecular interaction between the two Man$_9$ molecules ($-0.9\,\text{kcal}\,\text{mol}^{-1}$). **C.** Per-sugar difference in vdW contributions to binding Man$_9$ between the N- and C-terminal domains of MVL, taken from simulations of the individual domains on complex with the ligand. Positive values indicate that the C-terminal domain interacts more favorably. The data indicate that MVL$_C$ makes more favorable interactions with the D1 arm.

organism, its *in vitro* catalytic activity may not be physiologically-relevant. Based on the lectin definition discussed above, catalytic activity disqualifies MVL from being a *true* lectin. It has been proposed that MVL is an example of an evolutionary intermediate in the process of acquiring enzymatic activity [160]; however, this claim needs to be further explored, due to the limited number of known MVL homologues and lack of understanding of *M. viridis* biology. No ability to degrade high-mannose oligosaccharides has been detected for MVL. While the mechanism by which the protein degrades poly-GlcNAc sugars is not understood, mutagenic studies revealed that C-terminal domain residue Asp75 is crucial for catalysis; the N-terminal domain possesses no catalytic activity, likely because it has an Ala at the corresponding position 16 [160]. In this study, we carried out MD simulations of poly-GlcNAc oligosaccharides with MVL$_C$ as the receptor.

GlcNAc$_3$. The binding mode of GlcNAc$_3$ (chitotriose) to MVL was determined by Bewley and coworkers through NMR—intermolecular NOE and saturation transfer difference (STD) experiments demonstrated that the reducing GlcNAc A interacts tightest with the protein [160]. This binding mode places the *N*-acetyl group of the reducing-end sugar into a deep binding cavity on the MVL surface (as with high-mannose sugars, Figure 4-1A). However, subsequent visual inspection of the NMR-derived model identified suboptimal contacts between the protein and ligand, in contrast to both the X-ray structure of MVL, and the MD simulation ensembles. In particular, the reducing-end *N*-acetyl group was suboptimally positioned within its binding cavity on the MVL surface. Thus, instead of starting with the NMR model, we instead performed "mutagenesis" *in silico*. β-D-mannopyranose and β-D-glucopyranose are 2-epimers; starting from the trisaccharide Man$_1$, we modeled GlcNAc_3 by replacing the C2 hydroxyl of the terminal mannose with a hydrogen and the C2 hydrogen with acetamide. We term this binding mode GlcNAc_3-A.

The ligand remained in the extended conformation during the simulation, adopting glycosidic ϕ/ψ dihedral angles of $-74.1 \pm 6.6°/110.0 \pm 6.9°$, respectively, for the A–B linkage; and $-80.0 \pm 6.6°/112.1 \pm 8.7°$, respectively, for the B–C linkage. We observed that the N-acetyl group of ring C donated a hydrogen bond to the carbonyl oxygen of Trp96. The C2 hydroxyl of Man_1 also donates a hydrogen bond to this group, but the mannose needs to adopt a staggered conformation to do this. The N-acetyl group of GlcNAc is axial and can engage this interaction while remaining in the extended conformation.

We also constructed a model of GlcNAc_3 in complex with MVL, in which the reducing-end sugar is not coordinated by the protein. We term this binding mode GlcNAc_3-B. The binding of GlcNAc_3-A to MVL_C buries approximately 60 Å^2 of additional surface compared to Man_1, totaling -37.3 ± 3.6 kcal mol^{-1}. The alternative GlcNAc_3-B binding mode buries less surface area than GlcNAc_3-A, and makes significantly reduced interactions (-29.4 ± 2.9 kcal mol^{-1}). Both binding modes of GlcNAc_3 are unfavorable electrostatically, and the difference between the two modes arises from the lost favorable packing interactions when the reducing-end GlcNAc is not tightly associated with the protein.

Ambiguity of the GlcNAc_4 binding mode. STD NMR experiments demonstrated that for Man_2-A, Man_3, and GlcNAc_3 in complex with MVL, the N-acetyl group of ring A experiences the highest STD resonance enhancement compared to other carbohydrate groups [160] (Figure 4-5). This suggests that for all three ligands GlcNAc A interacts most closely with MVL, in agreement with the binding mode of high-mannose sugars. Crucially, for GlcNAc_4, *significantly more combined resonance enhancement was seen for rings B and C than for ring A* (Figure 4-5A). In addition, titration calorimetry binding experiments of GlcNAc_4 to MVL could not be analyzed using a single-site model (C. Bewley, personal communication).

Mechanistic studies following the degradation of GlcNAc_3 by wildtype MVL revealed that the protein cleaves the carbohydrate between rings B and C (Figure 4-5**A**). Similar experiments which monitored the degradation of GlcNAc_4 revealed the presence of the expected GlcNAc_2 cleavage product, as well as a small abundance of GlcNAc_3. We believe that the presence of GlcNAc_3 as a cleavage product of GlcNAc_4 is indicative of the heterogeneous binding mode of the substrate. In particular, since GlcNAc_4 is stable in solution indefinitely, possible binding modes of GlcNAc_4 to MVL may differ in their register (Figure 4-5B-C). We believe that these data argue in favor of altered or multiple GlcNAc_4 binding modes to MVL.

GlcNAc_4. We hypothesized that the two conformations of GlcNAc_4 binding to MVL differ in their register, and constructed two models of this complex (Figure 4-6). GlcNAc_4-A and GlcNAc_4-B extend GlcNAc_3 at the non-reducing and reducing ends, respectively. The GlcNAc_4-A model was rigid in simulation, with ring heavy-atom RMSD of 1.5 Å, and adopted the extended conformation exclusively. We observed more fluctuation at the non-reducing end of the ligand. A number of

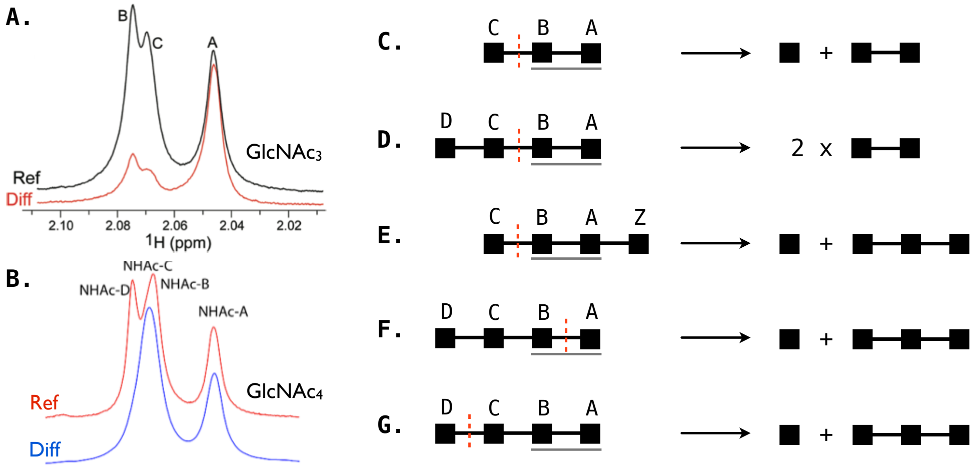


Figure 4-5: **Experimental evidence suggests heterogeneous binding by GlcNAc$_4$.** **A**. STD NMR spectrum of GlcNAc$_3$ in complex with MVL reveals that ring A experiences the most resonance enhancement. For clarity, only the *N*-acetyl portion of the spectrum is shown. **B**. STD NMR spectrum of GlcNAc$_4$ in complex with MVL reveals that GlcNAc residues B and C experience more combined resonance enhancement than GlcNAc A, suggesting that the binding mode of GlcNAc$_4$ is different than that of GlcNAc$_3$. Spectra and panels **A** and **B** were adapted from reference [160]. **C**. Schematic representation of *N*-acetyl-chitooligosaccharide cleavage by MVL. GlcNAc residues are represented by ■; the reducing-end residue is oriented to the right (ring A, unless Z is present). MVL cleaves chitotriose between rings B and C (dashed red line), producing GlcNAc$_2$ and GlcNAc. **D**. If GlcNAc$_4$ binds MVL in the same mode as GlcNAc$_3$, cleavage between rings B and C yields two GlcNAc$_2$ molecules, which do not bind MVL and cannot be processed further. This mode of binding is referred to as GlcNAc$_4$-A in the text. **E**. The presence of GlcNAc$_3$ as a cleavage product of GlcNAc$_4$ can be explained by an altered binding mode (referred to as GlcNAc$_4$-B in the text). In the register-shifted mode, cleavage between rings B and C yields GlcNAc$_3$ and GlcNAc; GlcNAc$_3$ is then processed as in **C**. **F–G**. An altered catalytic mechanism can also be invoked to explain the presence of GlcNAc$_3$. If GlcNAc$_4$ binds MVL in the same mode as shown in **D** and catalytic activity is directed toward a different glycosidic linkage (A–B or C–D), then GlcNAc$_3$ is produced in the reaction.

times during the simulation the ligand rotated out of the binding site as a rigid unit, anchored by the *N*-acetyl group of ring A. During the simulation of GlcNAc$_4$-B, the reducing-end GlcNAc Z adopted two alternate conformations; *k*-means clustering of the Z–A torsional angles revealed two clusters containing 22% and 78% of the population. The lower and higher populated clusters has ϕ/ψ values of $-54.1 \pm 9.4°/-34.2 \pm 11.0°$ and $-80.3 \pm 11.2°/102.7 \pm 17.8°$, respectively. The glycosidic linkages A–B and B–C explored the extended conformation exclusively.

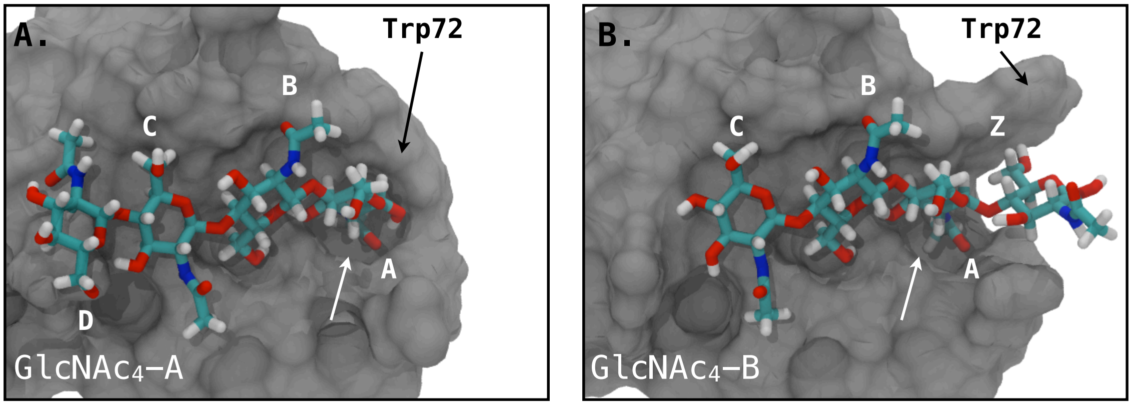


Figure 4-6: **Two models of GlcNAc$_4$ recognition by MVL.** The protein is shown in surface representation in both panels, and the sugar in licorice. The deep *N*-acetyl binding cavity is indicated with an arrow. The residue naming scheme in the two models consistently identifies the core GlcNAc$_2$ portion as A and B; the two ligand binding modes differ in their register. **A**. Model pose of GlcNAc$_4$ in which the reducing-end GlcNAc A *is* coordinated by the *N*-acetyl binding cavity. This model (GlcNAc$_4$-A) corresponds to extending GlcNAc$_3$ at the non-reducing end. **B**. Model pose of GlcNAc$_4$ in which the reducing-end GlcNAc Z is not coordinated by the *N*-acetyl binding cavity. This model (GlcNAc$_4$-B) corresponds to extending GlcNAc$_3$ at the reducing end. Note the altered conformation of Trp72 in the two binding modes.

The binding of GlcNAc$_4$ to MVL$_C$ buries approximately 870 Å^2 of SASA (Table 4.2) in both binding modes. We computed binding energies of -40.0 ± 7.4 and -38.4 ± 4.1 kcal mol^{-1}, for GlcNAc$_4$-A and GlcNAc$_4$-B, respectively. As with other ligands, the majority of the binding free energy (-38.6 and -36.8 kcal mol^{-1}, respectively) was derived from the nonpolar portion of the equation, with minor electrostatic contributions. The per-sugar breakdown of the van der Waals contributions are listed in Table 4.3.

GlcNAc$_5$. The insoluble polysaccharide chitin is composed of extended linear chains of GlcNAc residues. We constructed a model of GlcNAc$_5$ by simultaneously extending GlcNAc$_3$ at the reducing and non-reducing ends. The ligand remained bound during the course of the simulation, and was dynamic at both polymer ends. We calculated the binding free energy of -38.0 ± 8.0, with -41.0 kcal mol^{-1} derived from the nonpolar terms. As was observed for GlcNAc$_4$-B, the ligand was not perfectly complementary to the binding site of MVL. Despite burying more surface area than GlcNAc$_4$ in both binding modes, the predicted binding energy of GlcNAc$_5$ is equal to or marginally worse than that of $GlcNAc_4$.

$GlcNAc_2$. The $GlcNAc_2$ disaccharide core common to all ligands saccharides in this study (Figure 4-2) is likely necessary for binding MVL. However, it alone is not sufficient, since MVL does not bind $GlcNAc_2$ under typical experimental conditions (HSQC-monitored titration, MVL at ≈100 µM, up to twenty-fold molar excess of sugar) [16]. Presently, the smallest known carbohydrate ligand of MVL (in terms of number of monosaccharide units and molar mass) is $GlcNAc_3$. During the course of the 400 ns simulation, $GlcNAc_2$ remained bound to MVL_C, adopting a conformation identical to when it is part of larger saccharides ($-75.8 \pm 7.0°/111.2 \pm 7.2°$ ϕ/ψ angles). The binding of $GlcNAc_2$ to MVL_C buries a total of 576 Å^2 of SASA (Table 4.2), the smallest magnitude of any carbohydrate in this work. The calculated binding energy totaled -28.0 ± 3.8 kcal mol^{-1}, with the majority (-26.7 kcal mol^{-1}) derived from the nonpolar portion of the equation, and electrostatics contributing favorably at -1.3 kcal mol^{-1}.

4.4 Discussion & Conclusions

In this work, we investigated binding of the lectin MVL to a panel of carbohydrate ligands. We found that the computed binding energy of all ligands in this study is dominated by intermolecular van der Waals interactions, whose contribution scales linearly with ligand size and solvent-accessible surface area buried upon binding (Table 4.2). Unlike Cyanovirin-N ([57,58]), carbohydrate recognition by MVL is unfavorable in the electrostatic sense. Man_1, Man_2-A, and $GlcNAc_2$ were the only ligands in which the electrostatic contribution to binding is predicted to be favorable. In these cases, the receptors and ligands overcome their respective desolvation events upon association. Larger ligands (such as Man_9) experienced more unfavorable electrostatic interactions; such ligands bury significant amounts of surface area, but lack the interaction complementarity with MVL in order to overcome their desolvation penalties. This may be because these ligands are not the natural ligands for MVL within *Microcystis viridis.*

The experimentally-determined affinity of MVL for its cognate ligands is in the nanomolar range (approximately 9.5–12 kcal mol^{-1}). MM/PBSA approaches to calculate interaction free energies typically yield binding energetics orders of magnitude greater than observed experimentally [57,58]. As a result, these approaches are not intended for calculating absolute binding free energies, but instead can be applied to predict the relative binding of a series of ligands.

One of the main reasons for disagreement between experimental and computational binding free energy evaluation is the fact that the MM/PBSA approach not explicitly consider the entropic costs of binding. Macromolecular association is typically entropically opposed, since upon binding both receptor and ligand lose configurational degrees of freedom. Carbohydrates are flexible in solution, and contain numerous rotatable bonds; for example, the undecasaccharide Man_9 contains 22 glycosidic bond degrees of freedom. Interaction energies derived from MM/PBSA approaches are typically dominated by the magnitude of enthalpic interactions. For example, vdW

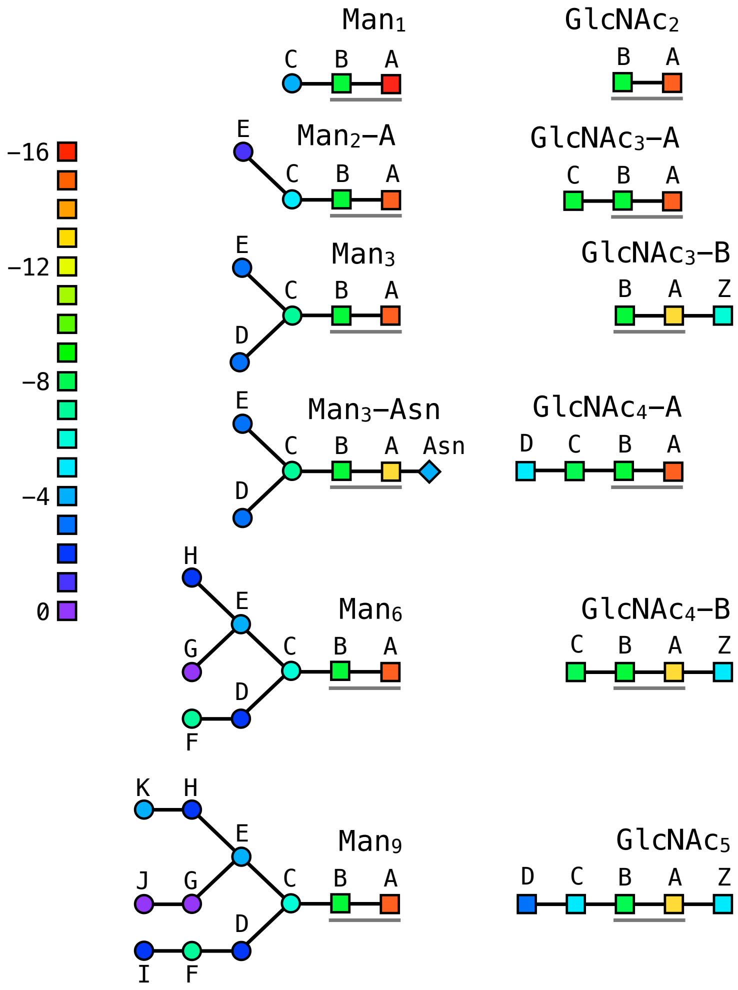


Figure 4-7: **Per-sugar decomposition of vdW interactions to binding.** Sugar residues are identified with the same symbols as Figure 4-2, and are color-coded by the magnitude of their vdW contribution to binding (purple to red, increasingly favorable). The energetic values are provided in Table 4.3. For all ligands, ring A makes the most significant contribution, and ring B the second-most significant. Extending the carbohydrate at the reducing end (as was done for Man$_3$-Asn, GlcNAc$_5$-A, or GlcNAc$_5$ weakens the interactions made by ring A, due to a re-arrangement of Trp72. In the case of high-mannose Man$_6$ and Man$_9$, the D2 arm comprising rings G and J makes to contributions to binding, compared to D1 and D3 arms.

interactions between a ligand a receptor scale proportionally with ligand size; for the series of high-mannose-derived ligands investigated in this study, we find that larger ligands are calculated as more favorable binders. To address the size-dependent bias, we need to consider the favorable dispersion interactions with solvent which both the receptor and ligand lose upon binding. The continuum van der Waals approach was designed to eliminate the source of size-dependent bias.

	Carbohydrate Model					
Residue	Man$_1$	Man$_2$-A	Man$_3$	Man$_3$-Asn	Man$_6$	Man$_9$
Asn				-4.2 ± 1.9		
A	-16.2 ± 1.8	-15.9 ± 1.8	-15.8 ± 1.9	-13.1 ± 2.6	-15.6 ± 1.9	-15.7 ± 1.8
B	-9.3 ± 1.4	-9.1 ± 1.4	-9.0 ± 1.3	-9.0 ± 1.4	-9.0 ± 1.3	-9.0 ± 1.3
C	-4.9 ± 1.9	-5.3 ± 2.3	-7.0 ± 1.3	-7.1 ± 1.2	-6.9 ± 1.5	-6.9 ± 1.3
D			-3.7 ± 1.2	-3.6 ± 1.2	-2.9 ± 1.1	-2.9 ± 1.1
E		-1.9 ± 1.6	-3.8 ± 1.2	-3.9 ± 1.1	-4.4 ± 1.0	-4.3 ± 1.0
F					-7.7 ± 2.4	-6.9 ± 2.6
G					-0.2 ± 0.2	-0.2 ± 0.2
H					-2.5 ± 0.9	-2.3 ± 0.8
I						-2.8 ± 1.7
J						-0.2 ± 0.3
K						-4.4 ± 1.5
Residue	GlcNAc$_2$	GlcNAc$_3$-A	GlcNAc$_3$-B	GlcNAc$_4$-A	GlcNAc$_4$-B	GlcNAc$_5$
Z			-5.9 ± 1.6		-5.0 ± 1.9	-5.3 ± 2.0
A	-15.8 ± 1.9	-15.9 ± 1.8	-12.8 ± 1.8	-15.7 ± 1.9	-13.3 ± 1.9	-13.4 ± 2.0
B	-9.0 ± 1.8	-9.7 ± 1.4	-8.5 ± 1.6	-9.3 ± 2.0	-9.6 ± 1.4	-8.9 ± 1.7
C		-9.4 ± 1.8		-8.3 ± 2.8	-8.8 ± 2.4	-5.8 ± 3.6
D				-5.5 ± 2.1		-3.9 ± 2.5

Table 4.3: **Per-sugar vdW contribution of carbohydrate models binding.** Contributions of high-mannose sugars (top half) and chitin-derived sugars (bottom half) are listed. All values are extracted from single-domain simulations of MVL$_C$, and are given in kcal mol^{-1}. The cumulative binding free energy $\Delta G^{\circ}_{\text{bind}}$, and the $\Delta G^{\circ}_{\text{vdW}}$ portion are given in Table 4.2. For all carbohydrate models in this study, GlcNAc A makes the strongest interaction with the protein, followed by GlcNAc_B.

Ligand	Receptor	$\Delta G^{\mathrm{rec}}_{\mathrm{cvdW}}$	$\Delta G^{\mathrm{lig}}_{\mathrm{cvdW}}$	$\Delta G^{\mathrm{binding}}_{\mathrm{cvdW}}$	ΔG_{vdW}	$\Delta G^{\mathrm{sum}}_{\mathrm{vdW}}$	$\Delta G^{\mathrm{corrected}}_{\mathrm{bind}}$
Man$_1$*	MVL$_\mathrm{N}$	+10.9	+15.9	+26.8	−30.4	−3.6	−7.9
	MVL$_\mathrm{N}$	+11.0	+15.8	+26.8	−30.6	−3.8	−8.0
	MVL$_\mathrm{C}$	+11.2	+16.2	+27.4	−30.2	−2.8	−8.4
	MVL$_\mathrm{C}$	+11.3	+16.2	+27.5	−30.7	−3.2	−8.1
Man$_2$-A	MVL$_\mathrm{N}$	+13.2	+17.2	+30.4	−33.4	−3.0	−5.9
	MVL$_\mathrm{C}$	+13.0	+17.2	+30.2	−31.8	−1.6	−5.8
Man$_3$	MVL$_\mathrm{N}$	+16.3	+19.3	+35.6	−38.6	−3.0	−4.9
	MVL$_\mathrm{C}$	+16.5	+19.7	+36.2	−39.1	−2.9	−5.5
	MVL$_\mathrm{C}$-E101R	+16.7	+20.3	+37.0	−40.3	−3.3	−6.5
Man$_3$-Asn	MVL$_\mathrm{C}$	+18.8	+21.2	+40.0	−41.9	−1.9	−4.3
Man$_6$	MVL$_\mathrm{N}$	+20.6	+23.9	+44.5	−46.2	−1.7	−2.9
	MVL$_\mathrm{C}$	+21.6	+24.7	+46.3	−49.0	−2.7	−5.4
Man$_9$	MVL$_\mathrm{N}$	+24.9	+26.8	+51.7	−50.6	+1.1	+1.2
	MVL$_\mathrm{C}$	+26.9	+28.6	+55.5	−56.0	−0.5	−1.3
GlcNAc$_2$	MVL$_\mathrm{C}$	+8.7	+13.9	+21.7	−24.7	−3.0	−6.3
GlcNAc$_3$-A	MVL$_\mathrm{C}$	+12.7	+17.0	+29.7	−34.9	−5.2	−7.6
GlcNAc$_4$-A	MVL$_\mathrm{C}$	+15.4	+19.3	+34.7	−38.6	−3.9	−5.3
GlcNAc$_4$-B	MVL$_\mathrm{C}$	+15.7	+18.7	+34.4	−36.8	−2.4	−4.0
GlcNAc$_5$	MVL$_\mathrm{C}$	+17.8	+20.3	+38.1	−37.2	+0.9	+0.1

Table 4.4: **Explicit consideration of loss of favorable dispersion interactions upon binding.** The continuum van der Waals approach tabulates the interactions lost by the receptor ($\Delta G^{\mathrm{rec}}_{\mathrm{cvdW}}$), ligand ($\Delta G^{\mathrm{lig}}_{\mathrm{cvdW}}$), and the complex ($\Delta G^{\mathrm{binding}}_{\mathrm{cvdW}}$) [8, 105]. Since these represent lost interactions, their sign is always positive; both the ligand and receptor contributions are directly proportional to solvent-accessible surface area lost upon binding. The magnitude of the solvent–complex and the molecular-mechanics tabulated intermolecular vdW interactions are similar, but opposite in sign. Incorporation of the continuum vdW model ($\Delta G^{\mathrm{sum}}_{\mathrm{cvdW}} = \Delta G^{\mathrm{binding}}_{\mathrm{cvdW}} + \Delta G_{\mathrm{vdW}}$) acts to correct the size-dependent bias in binding energetics.

Chapter 5

Biophysical and Functional Analyses Suggest that Higher-Order Multimetization is Required for the Activity of Adenoviral E4–ORF3*

Abstract

The early region 4 open reading frame 3 protein (E4–ORF3, UniProt ID: P04489) is the most highly conserved of all adenovirus-encoded gene products at the amino acid level. A key attribute of the E4–ORF3 proteins of different human adenoviruses is the ability to disrupt PML nuclear bodies from their normally punctate appearance into heterogeneous filamentous structures. This E4–ORF3 activity correlates with the inhibition of PML-mediated antiviral activity. The mechanism of E4–ORF3-mediated reorganization of PML nuclear bodies is unknown. Biophysical analysis of the recombinant WT E4–ORF3 protein revealed an ordered secondary/tertiary structure and the ability to form heterogeneous higher-order multimers in solution. Importantly, L103A, a non-functional E4–ORF3 mutant, forms a stable dimer with wild-type secondary structure content. Since the L103A mutant is incapable of PML reorganization, this result suggests that higher-order multimerization of E4–ORF3 may be required for the activity of the protein. In support of this hypothesis, we demonstrate that L103A acts as a dominant-negative effector when coexpressed with the WT E4–ORF3 in mammalian cells, presumably by binding to the WT protein and inhibiting the formation of higher-order multimers. *In vitro* protein binding studies support this conclusion. These results provide new insight into the properties of the adenovirus E4–ORF3 protein and suggest that higher-order protein multimerization is essential for E4–ORF3 activity.

*Portions of this chapter have been previously published as:

Patsalo V†, Yondola MA†, Luan B, Shoshani I, Kisker C, Green DF, Raleigh DP, Hearing P. Biophysical and Functional Analyses Suggest that Higher-Order Multimerization is Required for the Activity of the Adenovirus E4-ORF3 Protein. *Journal of Biological Chemistry*, 287(27):22573-83. (2012)

As part of this investigation, V. Patsalo purified recombinant E4–ORF3, performed CD and fluorescence spectroscopic experiments, AUC and DLS hydrodynamic characterization, mass spectrometry, and *E. coli* co-expression studies.

5.1 Introduction

PML nuclear bodies (PML-NB) are punctate multi-protein complexes linked to a variety of important cellular processes including transcriptional regulation, cellular growth control, DNA damage repair, apoptosis, and response to interferons (IFNs) [60]. A number of DNA and RNA viruses disrupt PML-NB organization upon virus infection to inhibit the antiviral activities of these structures. The mechanisms by which this is accomplished involve different processes, including proteasome-dependent degradation of PML (herpes simplex virus type 1) or relocalization/reorganization of PML (adenovirus (Ad) and human cytomegalovirus) [51]. Following Ad infection, PML-NB are relocalized within the nucleus by E4–ORF3 into heterogeneous filamentous structures termed "tracks" [43]. The E4–ORF3 protein recruits a number of cellular proteins into nuclear tracks (e.g., PML and Daxx) to inhibit their antiviral effects [176, 177]. E4–ORF3 also relocalizes two PML-NB-associated cellular transcription factors TIF1α and TIF1γ into tracks, perhaps related to the regulation of cellular gene expression [55, 180, 198]. Although E4–ORF3 shares a common localization with PML in the nuclear matrix fraction isolated from Ad-infected cells [30, 103, 104], the mechanism underlying PML-NB reorganization by E4–ORF3 is not known, nor is it understood how E4–ORF3 recruits other cellular proteins into these structures. The size (>1 µm) and appearance of the E4–ORF3-dependent tracks are consistent with the hypothesis that the E4–ORF3 protein multimerizes into higher-order structures.

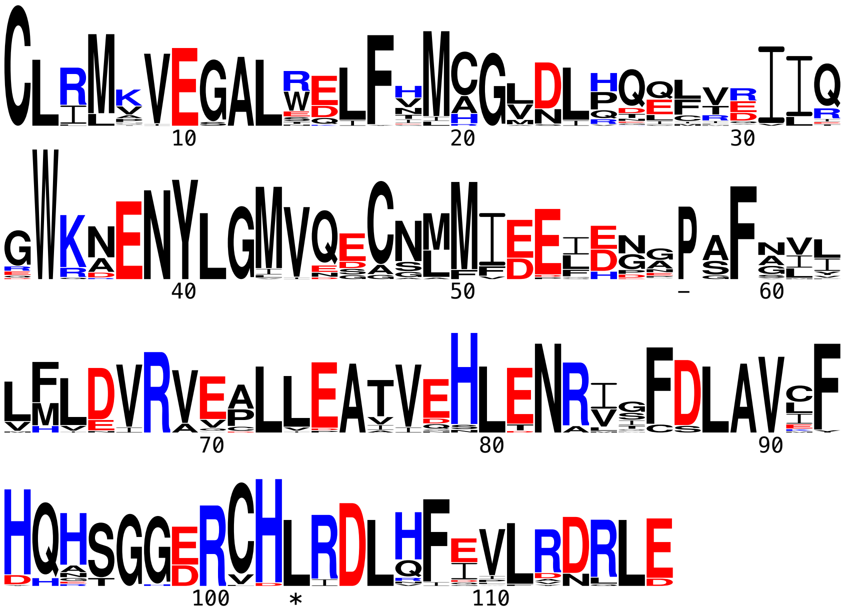


Figure 5-1: **Sequence logo representation of E4–ORF3 family.** The sequence logo is rendered with WebLogo 3.0 [37]; the height of the letter stack at each position is proportional to sequence conservation, while the height of the letters within a stack is scaled to relative amino-acid frequency. The numbers below the logo are residue numbers in WT E4–ORF3 sequence. Site of L103A mutation is indicated with $*$.

Viruses with linear dsDNA genomes such as Ad encounter a number of host cell responses that may severely inhibit viral replication [185]. The open ends of the linear viral genome are sensed by the host cell as broken DNA, triggering a cellular DNA damage response (DDR) [186]. The DDR severely inhibits Ad DNA replication if unabated [186] since the ends of the viral genome are ligated *via* the non-homologous end joining pathway; this results in the loss of DNA sequences at the multimeric junctions [89, 184] which contain the Ad origins of DNA replication. Additional sensors recognize dsDNA in the cytoplasm of infected cells early after viral infection and activate an IFN response [7]; this response can block virus replication by multiple mechanisms, including the inhibition of viral gene transcription and DNA replication. It is thus crucial for dsDNA viruses to counteract these antiviral host cell responses early after virus infection in order for a productive replication cycle to ensue.

Ad has evolved two redundant mechanisms to inhibit a cellular DDR [186]. The Ad5 E4–ORF3 protein relocalizes nuclear proteins involved in DDR, including Mre11, Rad50, and Nbs1 (the MRN complex) into the track structures [50, 173]. Available evidence suggests that E4–ORF3 inhibits the functions of these DNA repair proteins by a sequestration mechanism to block their access to the viral genome [50, 98, 123, 124, 173]. The Ad5 E1B-55K/E4-ORF6 complex functions as an adaptor molecule in an E3 ubiquitin ligase complex [72, 145]; E1B-55K/E4-ORF6 target the MRN proteins, as well as DNA ligase IV and Blm helicase, for inactivation via ubiquitin-mediated, proteasome-dependent degradation [5, 139, 173]. The E4–ORF3 protein facilitates this process by promoting the transport of the MRN proteins to cytoplasmic aggresomes [4, 111]. Either mechanism alone is sufficient to inhibit a DDR and allow efficient Ad DNA replication to occur. Finally, the E4–ORF3 protein inhibits p53-induced gene expression by establishing heterochromatin at p53-responsive cellular promoter regions [170]. The E4–ORF3 and E1B-55K/E4-ORF6 proteins are multifunctional and play additional roles in the viral life cycle including the promotion of cell cycle-independent viral replication, the regulation of viral late mRNA splicing and cytoplasmic mRNA accumulation, and the regulation of late protein translation [21, 62, 63, 136, 137, 150, 164, 165, 193]. Mechanisms underlying the roles of these Ad proteins in the aforementioned processes are poorly understood.

To probe the properties of the E4–ORF3 protein, we established a method to express soluble E4–ORF3 in *E. coli* and purify the protein utilizing, in part, its solubility properties in buffer containing sodium chloride. CD spectroscopy and coprecipitation experiments were used to demonstrate that the recombinant protein is both stably and correctly folded. We examined the *in vitro* biophysical properties of the WT E4–ORF3 protein and a non-functional mutant protein L103A [50]. Our results demonstrate that the WT E4–ORF3 protein forms heterogeneous multimers in solution. In contrast, L103A preferentially forms a stable dimer. Both proteins are similarly folded and possess equivalent secondary and tertiary structural characteristics, as demonstrated by CD and fluorescence emission spectroscopy. These data demonstrate that the loss of function of L103A is not due to improper folding and is consistent with the idea that the dimer formed by the L103A protein represents an intermediate in higher-order protein assembly.

We propose that the L103A mutant is nonfunctional because it is largely trapped in a dimeric state and deficient in higher-order multimerization required for track formation. A consequenct prediction of this mode is that the L103A mutant could dimerize with WT E4–ORF3 and prevent further multimerization.

In support of this hypothesis, expression of the L103A protein *in vivo* was found to block nuclear track formation and PML-NB reorganization by the WT E4–ORF3 protein and the WT and L103A protein were found to interact *in vitro*. In addition, the L103A mutant protein interfered with the interaction of WT E4–ORF3 with the cellular binding partner TIF1γ. These results provide new insight into the properties of the multifunctional E4–ORF3 protein and suggest that higher-order multimerization may be essential for E4–ORF3 activity.

5.2 Materials & Methods

Plasmid Constructs and Purification of E4–ORF3. The WT Ad5 E4–ORF3 reading frame was cloned by PCR into bacterial expression plasmids pET-15b (Novagen), pET-duet (Novagen), and pProEx-HTb (Invitrogen) while the non-functional E4–ORF3 mutant, L103A, was cloned into pProEx-HTb for protein expression in *E. coli*. These constructs expressed E4–ORF3 with a His$_6$-tag fused to the N-terminus and differed only in the protease cleavage sites used for removal of the His$_6$-tag. WT E4–ORF3 protein also was expressed using the IMPACT-CN system (New England Biolabs) where E4–ORF3 was expressed as a bipartite fusion protein containing a self-splicing intein and a chitin binding domain. Recombinant proteins expressed from each of these vectors behaved identically. The WT and L103A E4–ORF3 sequences fused to the His$_6$-tag and TEV protease cleavage site were excised from the pProEx-HTb vectors and inserted into mammalian expression plasmid pcDNA3 (Invitrogen).

Recombinant E4–ORF3 protein was expressed using the BL21 DE3 strain of *E. coli* (Novagen). Cultures were induced with 0.5 mM IPTG at an OD$_{600}$ = 0.6–0.8 after cooling to room temperature. Inductions were performed for 20 h at 25 °C and cell pellets were frozen prior to protein purification. In order to generate approximately 5 mg of pure WT E4–ORF3, cell pellets from a 2 L culture were lysed in 300 mL of lysis buffer (10 mM Tris-HCl, 10 mM βME, pH 7.5). E4–ORF3 contains three cysteine residues and reducing agent was included, as indicated, to prevent disulfide bond formation. Cells were lysed by sonication. Lysates were clarified by centrifugation at 20 kg for 30 min. The soluble fraction was decanted and NaCl was added to a final concentration of 100 mM. The sample was then agitated at 4 °C for 20 min. E4–ORF3 protein was precipitated by centrifugation at 10 kg for 15 min and the precipitate was resuspended in an equal volume of 10 mM Tris-HCl, 10 mM βME, pH 8.5 by mechanical disruption followed by agitation at 4 °C for 20 min. The sample was centrifuged at 13 kg for 20 min to remove any protein that did not resuspend. The soluble fraction was recirculated over a Ni^{2+}-NTA column overnight, the column was washed with 5 volumes of 10 mM Tris-HCl, 10 mM βME, 20 mM imidazole, pH 8.5. E4–ORF3 protein was eluted with the same buffer containing 200 mM imidazole. Individual fractions were assayed for protein by Bradford assay (Bio-Rad) and sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE). Size exclusion chromatography, the ultimate purification step was performed using a Superdex 200 16/60 column (GE Healthcare) in Buffer A (10 mM Tris-HCl, 0.5 mM TCEP, pH 8.5). A separate purification of the WT E4–ORF3 protein used a Superdex 75 26/60 column to take advantage of the increased loading capacity.

For *E. coli* co-expression studies, the gene for WT E4–ORF3 was cloned into pET-28a, where the protein product lacked a His_6 affinity tag. BL21 (DE3) cells were transformed with pProEx-L103A plasmid (Amp), and chemically-competent cell stocks were made. These cell stocks were then transformed with pET-28a-WT (Kan), and plated on LB plates containing both antibiotics. Protein expression was performed as indicated above, except that both ampicillin and kanamycin were included in the culture medium. The salting out purification step was skipped to avoid coprecipitating the two proteins.

***In Vitro* Binding Analyses.** Whole cell extracts of N52.E6 cells (a cell line that expresses the Ad5 E1A and E1B proteins [154]; and HeLa cells were prepared by cell lysis in F lysis buffer (10 mM Tris-HCl, 50 mM NaCl, 10% glycerol, 0.5% Triton-X100, 5 µM ZnCl_2, pH 7.5) on ice for 20 min followed by clarification for 30 min at 20 kg at 4 °C. 2 mg whole cell extract was incubated in 650 µL of F lysis buffer with E4–ORF3 WT or L103A protein at 7.5 µM on ice for 3 h, followed by the addition of anti-E1B-55K [107]; N52.E6 cell extract) or anti-TIF1γ [180]; HeLa cell extract) rabbit polyclonal antiserum and protein A-agarose beads. The samples were rotated at 4 °C overnight and then washed 5 times with F lysis buffer at 4 °C. The final protein A precipitates were boiled in SDS sample buffer containing 50 mM dithiothreitol for 10 min. Samples were separated by SDS-PAGE and analyzed by Western blot using anti-E1B-55K [107], anti-TIF1γ [180], or anti-E4–ORF3 [134] antibodies. Western blots were analyzed by enhanced chemiluminescence (Amersham Life Sciences) according to the manufacturers instructions. The same approach was used to characterize the binding of WT E4–ORF3 and the L103A mutant protein to TIF1γ following adenovirus infection. HeLa cells were infected with recombinant Ad5 E1-replacement vectors [49] that expressed either FLAG-tagged WT E4–ORF3 protein, HA-tagged L103A mutant E4–ORF3 protein, or with both viruses simultaneously, at the multiplicity of infection described in the text. Whole cells extracts were prepared 24 h after infection using F lysis buffer and sonication. WT E4–ORF3 was immunoprecipitated using anti-FLAG monoclonal antibody M2 (Sigma-Aldrich); TIF1γ was immunoprecitated using anti-TIF1γ antibody [180]. WT E4–ORF3 and L103A coimmunoprecipitation was analyzed using anti-E4–ORF3 [134] antibody, as described above.

Analytical Ultracentrifugation. Sedimentation equilibrium experiments were conducted at 25 °C at three rotor speeds (12000, 18000, 32000 rpm) on a Beckman Optima XL-A analytical ultracentrifuge. Typically, three samples at concentrations of 40, 20, 10 µM (A_{280} of 0.6, 0.3, 0.15) in Buffer A were prepared. The solute partial specific volume of 0.7403 cm^3/g was calculated based on its amino acid composition, and the solvent density of 0.9985 g/cm^3 was estimated from standard tables (10 mM Tris-HCl). Data were analyzed using HeteroAnalysis software (University of Connecticut Analytical Ultracentrifugation Facility). The data were fit globally across samples and speeds to yield the apparent molar mass.

Dynamic Light Scattering. Measurements were carried out using a Brookhaven Instruments 90Plus Particle Size Analyzer at a wavelength of 659 nm. For the L103A mutant protein, a sample of 400 µM protein concentration was used. For WT E4–ORF3, fractions A, B, and C eluting near the void volume of the Superdex 75 26/60 column were studied (see Section 5.3); the protein concentration of these samples was estimated to be 10, 15, and 10 µM, respectively (A_{280} of 0.142, 0.223, and 0.154). Ten data collection runs of 1 min–2 min each were acquired and combined; data were analyzed with the method of cumulants to yield a log-normal distribution of particle sizes, per manufacturer instructions.

Fluorescence Spectroscopy. Intrinsic tryptophan fluorescence spectra were collected at 25 °C on a PTI spectrofluorometer. The protein concentration was 12 µM (A_{280} of 0.18) in Buffer A. Fluorescence emission spectra (310 nm–400 nm) were collected upon selective excitation of the Trp fluorophore at 290 nm. Fluorescence excitation spectra (emission monitored at 350 nm) were collected upon excitation at a range of 250 nm–350 nm. Acrylamide quenching studies were carried out as a titration of increacing concentration of acrylamide dissolved in Buffer A; the final fluorescence intensity was corrected for dilution effects. The excitation wavelength was set to 295 nm. Data were normalized by the fluorescence intensity in the absence of quencher, and analyzed by linear regression to give the Stern–Volmer constant.

Circular Dichroism Spectroscopy. Measurements were carried out at 25 °C using a Chiroscan spectrometer (Applied Photophysics) equipped with a Peltier temperature controller. Far-ultraviolet spectra were collected from 190 nm–260 nm with sample concentrations of 10 µM and path length of 1 mm. Near-ultraviolet spectra were collected from 260 nm–340 nm on samples at 60 µm–100 µm concentration and path length of 10 mm In both cases, five spectra were collected (1 nm increments, 2 s nm^{-1}), averaged and smoothed using a Savitzky–Golay filter to yield the final result.

SAXS Experiments. Scattering experiments were performed at beamline X9 at Brookhaven National Laboratory, National Synchrotron Light Source I (Upton, New York, USA). Protein samples were injected to flow through a 1 mm-diameter capillary continuously during the measurement, at a rate of 0.67 µL s^{-1} in order to avoid radiation damage. The exposure time for each measurement was 30 s. Each sample was measured thrice and the results were averaged before data analysis. The program PRIMUS was used for buffer subtraction, and the radius of gyration (R_g) was obtained by the Guinier approximation: $I(q) = I(0) \cdot \exp(R_g^2 q^{2/3})$, where I is the intensity at scattering angle q.

***In Vivo* Analyses of E4–ORF3 Localization.** A pcDNA3 vector expressing WT E4–ORF3 fused to an N-terminal His$_6$-tag plus TEV protease cleavage site was transfected into HeLa cells

(ATCC) on glass coverslips for immunofluorescence (IF) assays using Nanojuice (Novagen) according to the manufacturers instructions. Transfected cells were washed with phosphate buffered saline (PBS) 24 h after transfection, fixed using 100% methanol at −20 °C for 5 min, washed with PBS, and blocked with PBS containing 10% goat serum for 1 h at room temperature. The coverslips were incubated with anti-E4–ORF3 rat monoclonal primary antibody [134] for 1 h at room temperature. The coverslips were then washed with PBS, incubated with FITC-conjugated anti-rat secondary antibody (Molecular Probes) for 45 min, washed with PBS, and the coverslips were mounted on slides using Immumount (Thermo Shandon). For E4–ORF3 localization analyses following virus infection, HeLa cells on glass coverslips were infected with recombinant Ad5 E1-replacement vectors [49] that expressed either FLAG epitope-tagged WT E4–ORF3 protein, HA epitope-tagged L103A mutant, or with both viruses simultaneously, at the multiplicity of infection described in the text. 24 h after infection, cells were processed for IF as described above using antibodies directed against the epitope-tags (rabbit anti-HA, Rockland Immunochemicals, and mouse anti-FLAG M2, Sigma-Aldrich), or anti-PML rabbit antibody (Santa Cruz Biotechnology) and anti-E4–ORF3 monoclonal antibody 6A11 [134], followed by the addition of FITC- and TRITC-conjugated secondary antibodies. Images were captured with a Zeiss digital deconvolution microscope equipped with Apotome and Axiovision 4.5 software.

5.3 Results

5.3.1 The Solubility of Recombinant E4–ORF3 is Affected by Ionic Strength and May be Exploited in Purification

Initial E4–ORF3 protein purification approaches utilized the IMPACT-CN system where the WT E4–ORF3 protein was expressed in *E. coli* as a bipartite fusion protein containing a self-splicing intein and a chitin binding domain. E4–ORF3 protein was largely insoluble when extracts were prepared using buffers that contained 500 mM NaCl. A significant portion of WT E4–ORF3 protein, however, was soluble in buffer containing only 20 mM Na-phosphate, 10 mM βME, pH 8.5. Intermediate levels of E4–ORF3 protein solubility were observed in buffers containing 50–500 mM NaCl (Figure 5-2).

Given these results, a novel purification strategy was implemented. WT E4–ORF3 was cloned into bacterial vectors for the expression of His$_6$-tagged protein. Recombinant E4–ORF3 expression was induced and *E. coli* were cultured overnight at 25 °C. *E. coli* extracts were prepared by sonication in buffer containing 10 mM Tris-HCl, 10 mM βME, pH 8.5. E4–ORF3 protein was salted out by the addition of NaCl to 100 mM and centrifugation. Upon resuspension in the original lysis buffer, a significant amount of the E4–ORF3 protein was resolubilized (Figure 5-2). Given the low salt conditions used for E4–ORF3 precipitation, very few bacterial proteins were precipitated. Once resuspended, the His$_6$-tagged protein was purified using a Ni^{2+}-NTA column (Figure 5-2) and subsequently by gel filtration. The solubility profile of the purified WT E4–ORF3 protein was similar to that observed with E4–ORF3 in *E. coli* cell extracts, although subtle differences were observed

(Figure 5-2). Addition of NaCl provided a simple and efficient purification step for E4–ORF3 proteins from *E. coli* cell extracts. However, the solubility profile of recombinant E4–ORF3 differed from that found when the protein was expressed in mammalian cells; WT E4–ORF3 protein expressed in HeLa cells was equally soluble in buffer containing 0–150 mM NaCl following sonication as observed by Western blot analysis. This may be due to lower levels of expression achieved in mammalian cells.

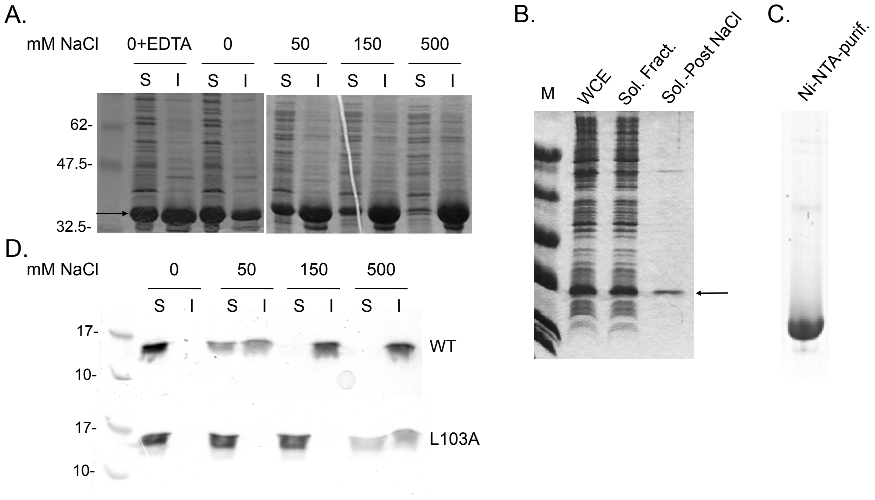


Figure 5-2: **Recombinant E4–ORF3 protein solubility and purification. A.** Cell pellets from 10 mL cultures of *E. coli* induced for the expression of recombinant WT E4–ORF3 protein were resuspended in lysis buffer containing 20 mM sodium phosphate, pH 7.5, and different concentrations of NaCl, as indicated in the figure. Cells were lysed by sonication and clarified by centrifugation prior to analysis by SDS-PAGE. The soluble fraction (S) was mixed 1:1 with SDS sample buffer and boiled. The insoluble fraction (I) was resuspended in SDS sample buffer and boiled. Equal amounts of the S and I fractions were analyzed by SDS-PAGE and coomassie blue staining. The E4–ORF3 protein is indicated by an arrow. **B.** Cell pellets from an *E. coli* culture induced for the expression of recombinant E4–ORF3 WT protein were extracted. Equal amounts of the whole cells extract (WCE), soluble fraction (Sol. Fract.), and soluble fraction following a salting out step (Sol.-Post NaCl) were analyzed by SDS-PAGE and Coomassie blue staining. The E4–ORF3 protein is indicated by an arrow. **C.** 6.5 µg of E4–ORF3 WT protein purified using a Ni^{2+}-NTA column (Ni-NTA-purif.) was analyzed by SDS-PAGE and coomassie blue staining. **D.** Purified WT E4–ORF3 and the L103A mutant protein were incubated at 20 µM in buffer, pH 7.5, containing different concentrations of NaCl, as indicated in the figure, on ice for 1 h.

5.3.2 His_6-TEV-E4–ORF3 is Functional *in Vivo* and *in Vitro*

During Ad infection, the WT E4–ORF3 protein forms filamentous nuclear structures referred to as tracks; these structures are heterogenous in an infected cell population [43]. To examine if the His_6-TEV tags present at the N-terminus of recombinant protein expressed in *E. coli* affected E4–ORF3 function in mammalian cells, the localization of untagged E4–ORF3 expressed during Ad5 infection was compared to that of His$_6$-TEV-tagged E4–ORF3 protein expressed by transfection. Both proteins displayed comparable nuclear tracks with the characteristic filamentous structures when examined by IF microscopy (Figure 5-9A–B). We conclude that an N-terminal His$_6$-TEV-tag on WT E4–ORF3 does not interfere with protein localization *in vivo.* In order to determine if recombinant WT E4–ORF3 was properly folded to allow for *bona fide* protein–protein interaction, coimmunoprecipitation assays were performed. Recombinant WT, and the non-functional mutant L103A [50] E4–ORF3 proteins were purified and added individually to cellular extracts prepared from HeLa cells or N52.E6 cells (as sources of the known E4–ORF3 binding partners TIF1γ [55, 180, 198] and Ad5 E1B-55K [134], respectively). TIF1γ and E1B-55K were immunoprecipitated using specific antibodies and coprecipitation of recombinant E4–ORF3 proteins was examined by Western blot (Figure 5-9C, IP). These data demonstrate that WT E4–ORF3 purified by these methods coprecipitates with its known binding partners TIF1γ and E1B-55K, indicating that the protein is properly folded. In contrast, the L103A mutant did not interact with TIF1γ and E1B-55K *in vitro.*

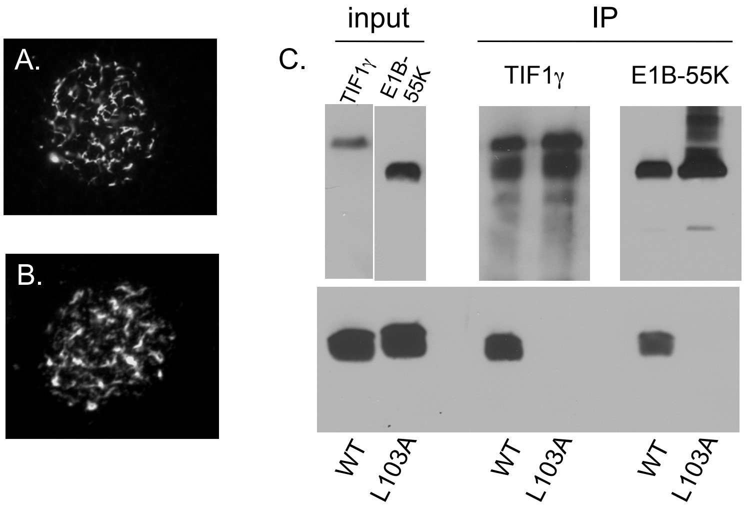


Figure 5-3: **His$_6$-TEV-tagged E4–ORF3 protein is functional. A.** Untagged WT E4–ORF3 was expressed by Ad5 infection of HeLa cells and subcellular localization analyzed by IF microscopy. The image is of one cell with E4–ORF3 tracks evident within the nucleus as filamentous structures. **B.** His$_6$-TEV-tagged E4–ORF3 was expressed by transient transfection of HeLa cells and subcellular localization analyzed by IF microscopy. **C.** The binding of E4–ORF3 WT and L103A proteins to TIF1γ and E1B-55K was analyzed by coimmunoprecipitation. Purified recombinant E4–ORF3 variants were added to cell extracts prepared from HeLa cells (for TIF1γ) or N52.E6 cells (for E1B-55K) and immunoprecipitations were performed using anti-TIF1γ- or anti-E1B-55K-specific antibodies. TIF1γ, E1B-55K, and the E4–ORF3 proteins were analyzed by Western blot using specific antibodies. Input, TIF1γ present in HeLa cell extract and E1B-55K present in N52.E6 cell extract (upper panels, left), and recombinant, purified E4–ORF3 WT and L103A proteins (lower panel, left). IP, immunoprecipitated TIF1γ and E1B-55K (upper panels, right) and coimmunoprecipitated E4–ORF3 WT and L103A proteins (lower panel, right).

5.3.3 Hydrodynamic Characterization of the E4–ORF3 protein

We noticed that purified L103A was less resistant to NaCl precipitation than WT E4–ORF3 (Figure 5-2). This observation, in addition to the inability of L103A to bind to TIF1γ and E1B-55K *in vitro*, suggested that the L103A mutation alters the biophysical properties of recombinant E4–ORF3 protein. To investigate this further, we turned to a detailed biophysical investigation of the WT and L103A E4–ORF3 proteins. Comparison of the recombinantly produced variants revealed that the altered functional properties of the two proteins *in vivo* may be attributable to a difference in their oligomerization state. We observed that the Superdex 200 size exclusion elution profiles of the two proteins differed significantly (Figure 5-4A). The WT protein elution profile was broad while L103A eluted as a well-separated and highly-abundant symmetric peak at 75 mL, suggesting that L103A exists as a monodisperse molecule. When chromatographed on a Superdex 75 column, the WT E4–ORF3 protein eluted as an asymmetric peak near the column void volume of 110 mL (Figure 5-4A–B). Both proteins eluted earlier than expected for a 16.2 kDa globular protein, suggesting oligomerization; the earlier elution times of WT E4–ORF3 suggested that it forms larger oligomerics than L103A. Three fractions from the WT E4–ORF3 near-void peak were chosen for further characterization (see below; fractions A, 112–116mL; B, 120–124mL; and C, 128–132mL; Figure 5-4B).

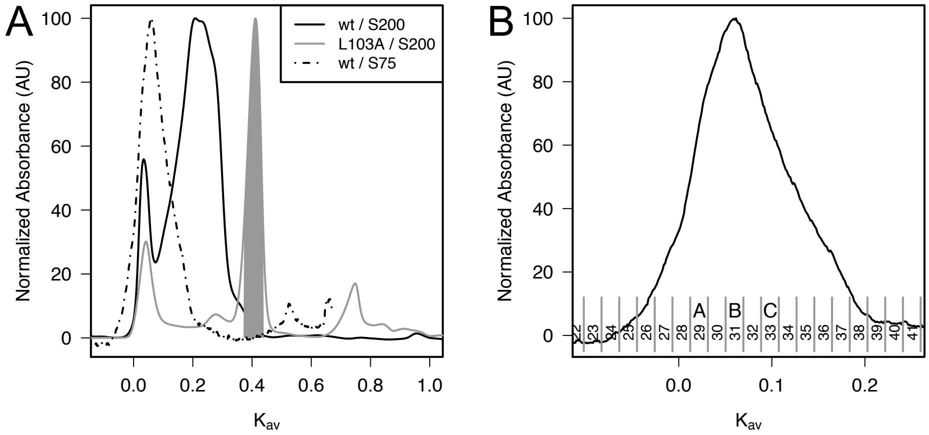


Figure 5-4: **L103A mutation alters E4–ORF3 quaternary structure.** **A**, Size exclusion chromatography was performed using Superdex 200 and Superdex 75 columns. The WT E4–ORF3 protein (solid black line) eluted as a complex mixture of oligomers. A large fraction of the L103A protein (gray line) eluted as a monodisperse peak (shaded) at K_{av} of 0.4 (≈75 mL) on a Superdex 200 16/60 column. This peak was collected and characterized for secondary structure content and oligomerization state. The elution times are normalized by column bed and void volumes: K_{av} of 0 corresponds to the column void volume. **B**, WT E4–ORF3 was fractionated using a Superdex 75 26/60 column (dotted line in panel **A**). Fraction numbers (22–41) and designations (A, B, and C) are indicated. The protein eluted as an asymmetric peak near the column void volume.

In order to test the hypothesis that recombinant WT and L103A E4–ORF3 proteins form oligomers, we set out to quantitatively evaluate their association state. Sedimentation equilibrium analytical ultracentrifugation (seAUC) experiments showed that the WT E4–ORF3 protein is heterogeneous in solution (Figure 5-5). The equilibrium heterogeneity is consistent with the size exclusion profiles of the protein eluting from both Superdex 200 and Superdex 75 columns (Figure 5-4A). Attempts to deconvolve the sedimentation profiles using an equilibrium model of discrete oligomeric assembly (monomer/N-mer or monomer/M-mer/N-mer) were not successful.

In contrast to WT E4–ORF3, the equilibrium sedimentation of L103A was well described as an ideal solute with apparent molecular weight of 30.3 kDa, as judged by the randomness of the residuals and the linearity of the transformed data (Figure 5-5A). The theoretical mass of the protein calculated from its sequence is 16.2 kDa; the AUC-estimated molecular weight is within 7% of a homodimer (32.4 kDa), within the experimental error of our instrument. This indicated that L103A exists as a dimer at µM concentrations. In addition, when L103A was chromatographed at high concentrations, we noticed an appearance of a left shoulder to the dimeric peak. Using seAUC, we found that fractions in this shoulder were well-fit to an ideal solute model with molecular weights corresponding to dimer-of-dimers, and trimer-of-dimers species.

While size exclusion chromatography can be used to estimate the size of proteins, ionic interactions with the column affect solute progress unless a significantly high salt concentration is included to screen them. The salt sensitivity of E4–ORF3 prevents this. As an alternative, light scattering techniques provide a way to estimate the size of particles in solution, and are less sensitive to ionic strength effects. In order to test the hypothesis that WT E4–ORF3 forms larger oligomers than L103A, we analyzed three fractions of WT E4–ORF3 (fractions A, B, and C; Figure 5-4B) and the dimeric L103A sample (Figure 5-4A) by dynamic light scattering (DLS). In this technique, the quasielastic light scattering intensity has a direct dependence on particle size. As a consequence, larger particles can be studied at significantly lower concentrations than smaller particles.

We found that it was necessary to concentrate dimeric L103A samples in order to achieve adequate light scattering intensity. The polydispersity increased slightly when the L103A mutant was concentrated, but remained a small percentage of the total signal (confirmed by sedimentation equilibrium AUC, data not shown). The effective hydrodynamic diameter of L103A was found to be 4.4 nm. In contrast to L103A, fractions of WT E4–ORF3 could be characterized at low µM concentrations. DLS revealed that the WT E4–ORF3 protein formed particles significantly larger than those present in the L103A sample (Figure 5-5B). Effective hydrodynamic diameters of 62.6, 18.2, and 5.9 nm were estimated for the the three WT E4–ORF3 fractions. It is important to note that smaller hydrodynamic diameter was observed for the L103A protein even at a 40-fold molar excess over the WT protein. These data demonstrate that the L103A dimeric state can not simply be overcome by concentration effects.

In addition, we used small angle X-ray scattering (SAXS) to examine the solution conformation of WT and L103A proteins. SAXS is sensitive to the overall shape and size of a molecule. The scattering profiles collected for the two proteins differed significantly (Figure 5-5). Inspection of the

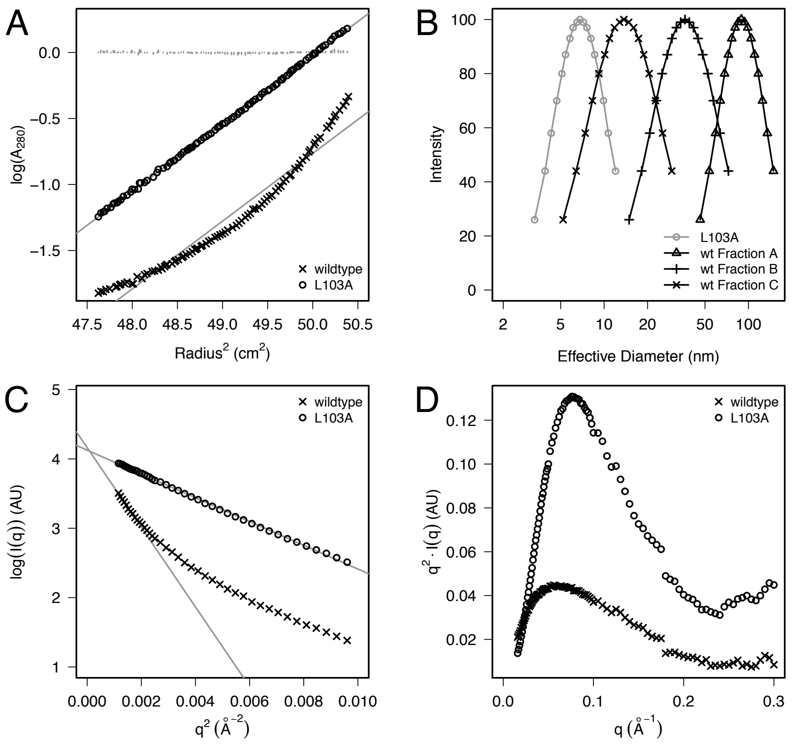


Figure 5-5: **Hydrodynamic characterization of WT and L103A E4–ORF3. A.** Sedimentation equilibrium data for the two variants. Molecular heterogeneity is evidenced by curvature of the sedimentation profile when log(A_{280}) is plotted versus squared radius. A least-squares regression line (gray) is displayed for each dataset; the residuals for the L103A protein are plotted in gray near $y = 0$. **B.** Dynamic light scattering properties of the E4–ORF3 WT (fractions A, B, and C; Figure 5-4B) and L103A proteins. Distribution function intensities versus effective particle diameter, derived from dynamic light scattering, indicate that WT E4–ORF3 assembles into larger oligomers than L103A. Note the logarithmic spacing of the x-axis. **C.** Guinier analysis of SAXS of WT and L103A. Least-squares linear regression fits of the linear portions of the two data sets are shown as solid grey lines. **D.** Kratky plots of the data in panel **C** show a peak in their profiles, indicating globular structures for both WT and L103A E4–ORF3.

small-angle region using the Guinier representation indicated molecular heterogeneity for the WT protein, as evidenced by data curvature. L103A protein produced a linear profile in the small-angle region, from which we extracted the radius of gyration (R_g) of 22.8 Å (2.2 nm) using the Guinier approximation.

The determined R_g is within the range observed for other proteins of similar length. For example, α-subunit tryptophan synthase is 268 residues long and has an R_g of 19.1 Å, while ospA has an R_g of 25.0 Å and is 257 residues long. A homodimer of E4–ORF3 contains 272 amino-acid residues, including the unstructured residues in the His_6-tag. The predicted R_g for a 136-residue folded monomer is 14.2 Å, while the value predicted for a fully unfolded 136-residue monomer is 36.4 Å. The predicted values are based on correlations between the length of a protein chain in residues and the observed R_g. The empirical relationship predicts an R_g of 18.5 Å for a dimer of 272 residues. The somewhat larger value observed for the L103A dimer likely reflects shape effects. The key point is that the experimental R_g is significantly larger than expected for a folded monomer and less than predicted for a fully unfolded monomer. The Kratky representation of the scattering data revealed the presence of a peak for both WT and L103A proteins, as is expected for a folded globular protein (Figure 5-5D). Thus, the SAXS studies are fully consistent with the AUC data and the gel filtration experiments.

5.3.4 WT E4–ORF3 is Well-Folded and L103A Adopts the Same Structure

In order to confirm that the protein fold is not perturbed by the L103A substitution, we used CD spectroscopy to probe protein secondary structure. Far-UV spectra of the E4–ORF3 WT and L103A proteins revealed that the two have identical secondary structure, with pronounced spectral minima at 208 and 222 nm and a maximum at 195 nm (Figure 5-6A); the spectra indicate significant helical content. These data demonstrate that the secondary structure is not perturbed by the L103A mutation. In addition, different gel filtration fractions of WT E4–ORF3 possessed identical secondary structure (Figure 5-7), suggesting that multimerization of E4–ORF3 is not accompanied by significant gain or loss of helical content.

E4–ORF3 contains a single Trp residue (Trp35), which we hypothesized may serve as a probe of tertiary structure in the protein. Near-UV CD spectra of WT and L103A variants (Figure 5-6B) revealed a number of narrow concordant bands arising from residues which absorb in this region (Trp, Tyr, Phe). A structured non-zero near-UV CD spectrum is a hallmark of a compact cooperatively-folded protein, and many partially-folded proteins give no near-UV CD signal even when the far-UV CD shows significant secondary structure. The spectral bands seen for WT and L103A were not identical in intensity, but displayed remarkable agreement in their position and line width. We further probed the environment of Trp35 using fluorescence spectroscopy. The fluorescence emission and excitation properties of Trp residues are sensitive to solvent accessibility and proximity to quenching groups. Figure 5-8 displays the fluorescence emission and excitation spectra of the E4–ORF3 WT and L103A proteins. The excitation spectra are superimposeable, indicating that the fluorophore is present at the same concentration in both samples, and that the

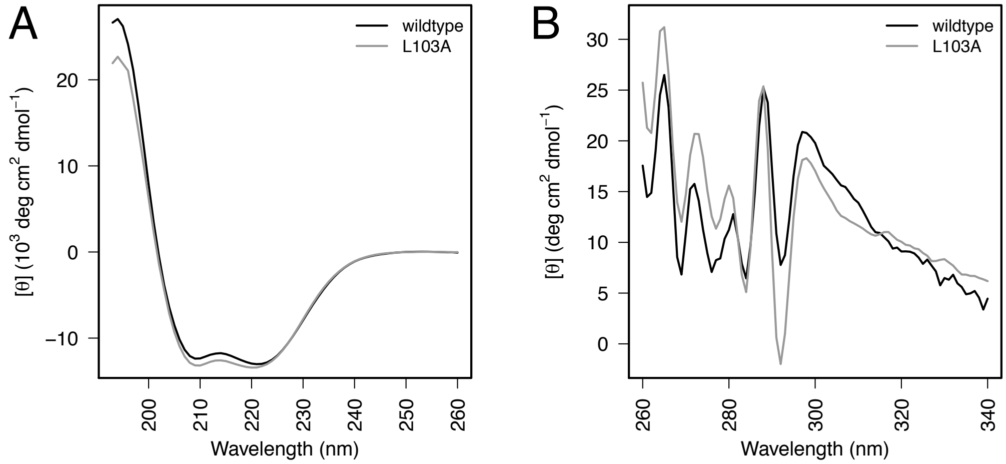


Figure 5-6: **CD characterization of E4–ORF3 variants. A.** Far-ultraviolet CD spectra of the E4–ORF3 WT and L103A denonstrate that the proteins contain nearly-identical secondary structure. Spectral minima at 208 and 222 nm, as well as the maximum at 195 nm suggest significant helical content of the proteins. **B.** Near-UV CD spectra collected for WT and L103A of E4–ORF3 revealed the presence of narrow bands for both variants, indicating a similar environment for the protein aromatic sidechains. The two variants contained concordant bands arising from dipole orientation and electronic structure of residues which absorb in this region (Trp, Tyr, Phe).

local fluorophore environments are similar.

When Trp35 was selectively excited, the emission spectra of the two variants indicate that the molecular environment is similar in both proteins. The emission maxima of the E4–ORF3 WT and L103A proteins were 326 and 323 nm, respectively; both were blue-shifted from a solvent-exposed maximum of approximately 350 nm. The quantum yield of L103A was approximately 10% greater than that of WT E4–ORF3. We used acrylamide quenching as an independent confirmation of solvent accessibility of Trp35. Fluorescence quenching of Trp by acrylamide requires the two to come into physical contact—the ease of quenching thus reports on the solvent accessibility of the indole side chain of Trp. We found that the ease of quenching the fluorescence of Trp35 in WT and L103A proteins was comparable, with a Stern–Volmer constant of approximately 1 M^{-1} (Figure 5-8B). The aforementioned results of three independent techniques agree, indicating that Trp35 is sequestered from solvent in both the WT and L103A E4–ORF3 proteins; the small change in the quantum yield could arise from subtle changes in the environment, but the overall environments are similar.

5.3.5 The L103A Mutant Inhibits the WT E4–ORF3 *in Vivo*

If the dimer formed by the E4–ORF3 L103A mutant protein represents an intermediate normally involved in higher order E4–ORF3 protein assembly, but is trapped by the mutation, then one would anticipate that L103A could interfere with higher assembly of the WT E4–ORF3 protein

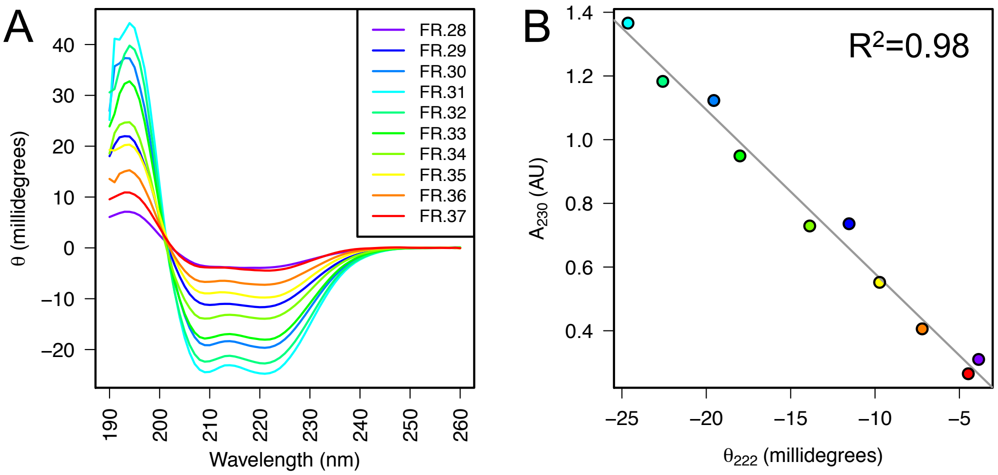


Figure 5-7: **CD demonstrates that secondary structure content of WT E4–ORF3 is multimerization-independent.** **A.** Far-UV CD spectra of size exclusion fractions of WT E4–ORF3 chromatographed on a Superdex 75 26/20 column. The spectra are given as raw measured ellipticity (are not concentration-normalized). **B.** Helical content (ellipticity at 222 nm) is presented as a function of protein concentration, given by A_{230}; each point represents a distinct sample for which we measured A_{230} and collected a CD spectrum. The colors used correspond to those in panel **A**. Helical content is directly proportional to protein concentration, indicating that all fractions possess similar secondary structure, and that the secondary structure of recombinant E4–ORF3 is not significantly altered by oligomeric state. A linear least-squares fit is shown as a gray line.

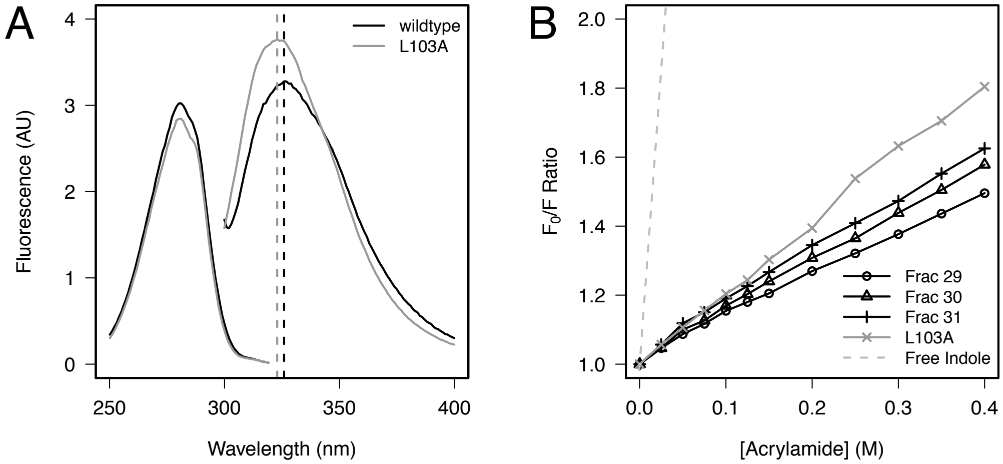


Figure 5-8: **Fluorescence characterization of E4–ORF3 variants. A.** Fluorescence excitation and emission spectra of WT and L103A suggest that Trp35, the unique Trp in the E4–ORF3 sequence, is partially sequestered from solvent and is in a similar environment in both proteins. Wavelength of maximal emission intensity of both samples is indicated wth a dashed vertical line. **B.** Collisional quenching of Trp fluorescence (F_0/F) by acrylamide is consistent with steady-state emission data. Stern–Volmer constants of approximately $1\,\text{M}^{-1}$ were derived both several samples of WT E4–ORF3 and for L103A. The dashed line indicates quenching of an indole, representing maximal solvent accessibility; the Stern–Volmer constant of quenching the indole is $14\,\text{M}^{-1}$.

by sequestering WT E4–ORF3 into nonproductive dimers. To test this hypothesis, recombinant Ad5 vectors were used that express WT and L103A proteins with separate epitope tags. The WT E4–ORF3 protein was tagged with the FLAG epitope and the L103A mutant protein was tagged with the HA epitope to allow the visualization of each protein within the same infected cell. HeLa cells were infected individually with Ad5 vectors that express each E4–ORF3 protein at low multiplicities of infection (MOI; 200 physical virus particles per cell corresponding to ≈10 infectious viruses per cell). Typical nuclear tracks were found in cells that express WT E4–ORF3 (Figure 5-9A), whereas diffuse nuclear localization of the L103A mutant protein was observed (Figure 5-9B), as previously reported [50]. Coinfections were performed with a low MOI (200 virus particles per cell) of the FLAG-tagged WT E4–ORF3 expression vector and a high MOI (1250 virus particles per cell) of the HA-tagged L103A expression vector. Our rationale was that over-expression of L103A relative to WT E4–ORF3 may drive the WT to preferentially form dimers with the L103A mutant. A significant reduction in the ability of the WT E4–ORF3 protein to form nuclear tracks was observed in coinfected cells. Representative examples are shown in Figure 5-9 (panels C–H) where two patterns of WT E4–ORF3 expression were detected. In some cells, small nuclear punctæ of WT E4–ORF3 that colocalized with the L103A mutant were evident (panels F–H). Of note, such punctæ were never observed with the L103A mutant protein alone at any MOI; therefore, the WT protein contributes to this process. In other cells, few nuclear punctæ of WT E4–ORF3 were observed; instead, a diffuse nuclear staining pattern of both WT and the L103A mutant protein was detected (panels C–E). These results were specific to the coexpression of the E4–ORF3 WT and L103A proteins since an augmentation of nuclear track formation was observed when cells were coinfected with a low MOI of FLAG-tagged WT E4–ORF3 vector and a high MOI of HA-tagged WT E4–ORF3 vector (panels I–K).

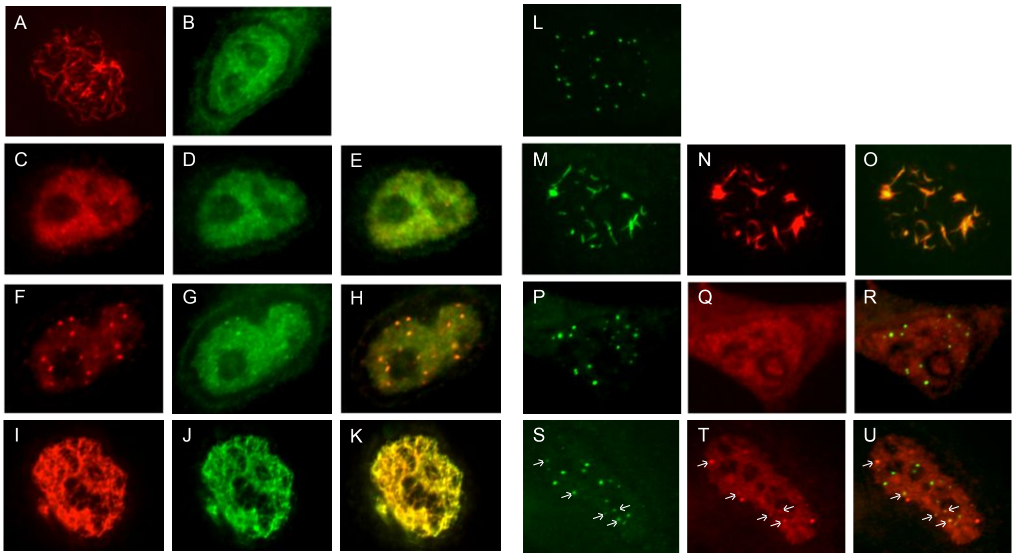


Figure 5-9: **L103A inhibits WT E4–ORF3 protein activity *in vivo*.** **A.** FLAG-tagged WT E4–ORF3 localization following infection of HeLa cells at a low MOI was analyzed by IF microscopy using TRITC-labeled secondary antibody. The image is of one cell with E4–ORF3 tracks evident within the nucleus as filamentous structures. **B.** HA-tagged L103A localization following infection of HeLa cells at high MOI was analyzed using a FITC-labeled secondary antibody. **C–H.** Coinfection at a low MOI of FLAG-E4–ORF3-WT plus a high MOI of HA-L103A. E4–ORF3 localization was analyzed by IF microscopy using FITC-labeled (anti-HA, panels **C** and **F**) and TRITC-labeled (anti-FLAG, panels **D** and **G**) secondary antibodies. Merged images are shown in panels **E** and **H**. **I–K.** Coinfection at a low MOI of FLAG-E4–ORF3 WT plus a high MOI of HA-E4–ORF3-WT. A merged image is shown in panel **K**. **L**, PML-NB evident in uninfected HeLa cells represented as discrete punctæ. **M–O.** PML-NB reorganization following infection of HeLa cells at low MOI with Ad-CMV-FLAG-WT E4–ORF3. **P-R.** PML-NB formation following infection of HeLa cells at high MOI with Ad-CMV-HA-L103A. **S–U.** PML-NB formation following coinfection of HeLa cells at low MOI with Ad-CMV-FLAG WT E4–ORF3 and high MOI Ad-CMV-HA-L103A. PML detected in panels **M**, **P**, and **S** using rabbit anti-PML antibody and FITC-conjugated secondary antibody. E4–ORF3 proteins were detected in panels **N**, **Q**, and **T** using anti-E4–ORF3 monoclonal antibody and TRITC- conjugated secondary antibody. Merged images are shown in panels **O**, **R**, and **U**.

The aforementioned results suggest that L103A may act as a dominant-negative effector to antagonize the activity of WT E4–ORF3. To analyze this possibility, we examined if L103A interferes with the ability of WT E4–ORF3 to reorganize PML-NB. HeLa cells were infected individually with Ad vectors that express FLAG-tagged WT E4–ORF3, HA-tagged L103A, or coinfected with both viruses. Subsequently, PML-NB relocalization was analyzed by IF microscopy (Figure 5-9). In uninfected cells, typical PML-NB were evident (panel L); PML was relocalized into nuclear tracks following expression of WT E4–ORF3 (panels M–O) but not L103A (panels P–R). Interestingly, L103A interfered with the ability of WT E4–ORF3 to relocalize PML (panels S–U). In coinfected cells, colocalization with PML-NB was observed with some, but not all, of the E4–ORF3 punctæ.

To further probe the interactions between WT and L103A E4–ORF3, *in vitro* binding studies were performed. We analyzed if the WT E4–ORF3 and L103A proteins directly interact by coexpressing the two in *E. coli* and in mammalian cells. In *E. coli*, untagged WT E4–ORF3 and His_6-tagged L103A were purified on a Ni^{2+}-NTA affinity column, followed by gel filtration. The salting out step was omitted to avoid coprecipitating the two proteins; we wanted to visualize protein complexes formed between the two during protein expression, rather than those that might arise as a consequence of salt precipitation. The gel filtration profile of the Ni-NTA purified proteins differed significantly from those of WT E4–ORF3 or L103A alone (Figure 5-10A), and contained a broad distribution of E4–ORF3 oligomers. Remarkably, we found that both proteins were present in each gel filtration fraction. SDS-PAGE revealed that WT E4–ORF3 preferentially populated larger oligomers than L103A and that fractions containing the highest abundance of L103A eluted at the same time as dimeric L103A characterized above. The identity of the two proteins was confirmed using MALDI-TOF mass spectrometry, and their abundance was estimated using SDS-PAGE and analytical HPLC (Figure 5-10B–C).

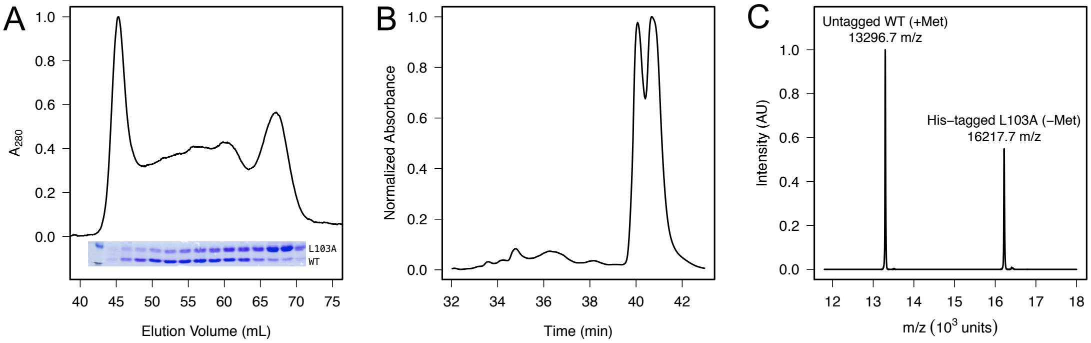


Figure 5-10: **Recombinant co-expression of WT and L103A variants of E4–ORF3. A.** Superdex 200 gel filtration profile of the Ni-affinity purified co-expression is shown. Individual fractions were probed for protein content by SDS-PAGE (overlaid.) Material eluting near the void volume of 45 mL was determined to be contaminating nucleic acids using SDS-PAGE and ethidium bromine fluorescence. Note the presence of both proteins in all examined gel filtration fractions. **B.** Gel filtration fractions from panel **A** were pooled and loaded onto an analytical C_4 HPLC column. The proteins were eluted with a linear gradient of 0–90% isopropanol, 0.1% trifluoroacetic acid. While the peaks were not resolved enough for accurate quantification, we approximate (using HPLC and SDS-PAGE from panel **A**) that a nearly-equal amount of both proteins was present. **C**, MALDI-TOF mass spectrometry was used to confirm the presence and identity of both proteins. For His_6-tagged L103A, the calculated MW is 16 221.5 Da (m/z of 16217.7 observed), after removal of N-terminal leading methionine. For untagged WT E4–ORF3, the calculated MW is 13 298.4 Da (m/z of 13296.7 observed); for this protein product, the N-terminal methionine was retained.

To investigate interactions between the two E4–ORF3 proteins in mammalian cells, HeLa cells were infected individually with Ad vectors that express FLAG-tagged WT E4–ORF3, HA-tagged L103A, or coinfected with both viruses. Subsequently, whole cell extracts were prepared and anti-FLAG antibody was used to immunoprecipitate WT E4–ORF3. Coprecipitation of the L103A mutant with WT E4–ORF3 was confirmed by Western blot (Figure 5-11 IP: FLAG).

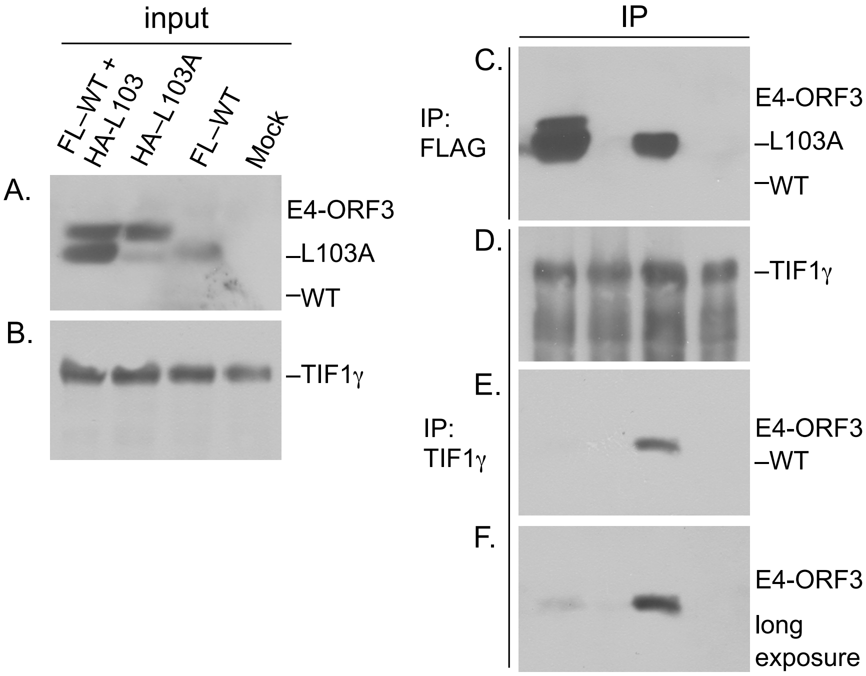


Figure 5-11: **WT E4–ORF3 and L103A coimmunoprecipitate and L103A inhibits binding of WT-ORF3 to TIF1γ.** HeLa cells were mock-infected (mock), individually infected with low MOI and high MOI of Ad-CMV vectors that express WT E4–ORF3 and HA-L103A, or coinfected with low MOI FLAG-WT-ORF3 and high MOI HA-L103A. E4–ORF3 (**A**) and TIF1γ (**B**) in cells extracts (input) were detected by Western blot. FLAG-WT-ORF3 migrated faster than L103A in the gel due to the presence of a smaller epitope tag; both proteins were detected using an anti-E4–ORF3 monoclonal antibody. The major L103A species corresponds to the epitope-tagged protein recognized using an anti-HA antibody; the weaker, faster migrating L103A was not recognized by an anti-HA antibody an may reflect translation initiation at an internal methionine residue or proteolytic cleavage of the epitope tag. E4–ORF3 (**C**, **E**, and **F**) and TIF1γ (**D**) proteins present in anti-FLAG (**C**) and anti-TIF1γ (**D**, **E**, and **F**) immunoprecipitates (IP) were detected by Western blot. Panel **F** is a longer exposure of the same gel shown in **E**.

Finally, we examined if the L103A mutant interferes with the ability of WT E4–ORF3 to bind to the known cellular binding partner TIF1γ. Single virus infections and coinfections were performed as described above, TIF1γ was immunoprecipitated from whole cell extracts, and E4–ORF3 binding analyzed by Western blot (Figure 5-11). These results demonstrated that the L103A mutant protein interfered with the binding of WT E4–ORF3 to TIF1γ (IP: TIF1γ). Collectively, these results are consistent with the hypothesis that the L103A mutant protein is able to sequester the WT E4–ORF3 protein within the nucleus and inhibit higher-order protein assembly. These results also support the conclusion that the stable dimer observed with the L103A mutant protein represents a bona fide intermediate in E4–ORF3 nuclear track formation.

5.4 Discussion & Conclusions

Our biophysical analysis shows that WT E4–ORF3 and the L103A mutant have identical secondary structure content and that both form compact globular structures with spectroscopic properties expected for native proteins. The major difference between the two proteins is their ability to oligomerize. DLS, seAUC, gel filtration, and SAXS all reveal the wildtype forms heterogeneous mixtures of high-order oligomers while self-assembly of the mutant is largely stopped at the dimer stage. Spectroscopic studies show that the wildtype oligomers contain native-like secondary structure and are folded. Thus, WT E4–ORF3 oligomerization fundamentally differs from non-specific aggregation of unfolded proteins.

Given the small size of the E4–ORF3 protein (116 amino acids) and large number of known cellular and viral interaction partners, we believe that higher-order multimerization is necessary for E4–ORF3-mediated pml-nb disruption. Higher order assembly will lead to the presentation of multiple binding sites. Our results are consistent with the hypothesis that multimeric WT E4–ORF3 has a higher affinity than L103A for cellular binding partners such as TIF1γ due to the presence of multiple binding interfaces on the WT protein. We propose that higher-order multimerization of the WT E4–ORF3 protein forms a polyvalent scaffold which is required for the activity of E4–ORF3. Each individual scaffold building block (a dimer) has low affinity for E4–ORF3 binding partners. However, this low affinity can be overcome through avidity effects; higher-order multimerization of E4–ORF3 in the nucleus presents a polyvalent surface which greatly reduces the entropic costs of binding. The fact that the dimeric L103A mutant protein is incapable of binding both the E1B-55K and TIF1γ proteins supports this hypothesis; L103A may contain the binding regions for E1B-55K and TIF1γ, but its inability to multimerize prevents it from raising this binding affinity to levels sufficient for efficient binding partner interactions.

This model is also supported by the results of IF experiments which demonstrated that coexpression of L103A with WT E4–ORF3 interfered with the ability of the WT protein to form nuclear tracks. L103A also interfered with the binding of WT E4–ORF3 to TIF1γ in coimmunoprecipitation experiments.

We hypothesize that upon coexpression, L103A functions as a dominant-negative effector by forming trapped heterodimers, or short multimers, with WT E4–ORF3. This hypothesis is supported by IF analyses which demonstrated the recruitment of L103A into small punctæ only upon coexpression with the WT protein. Also supporting this hypothesis are the observations that WT E4–ORF3 and L103A coimmunoprecipitate from cellular extracts when expressed *in vivo* and cofractionate from cellular extracts when coexpressed in *E. coli.* While we favor the possibility that WT E4–ORF3 and the L103A mutant heterodimerize resulting in interference of WT activity by L103A, it is also possible that the region near L103 may identify a critical binding surface for cellular binding partners which is perturbed upon mutation and that such interactions are required for higher order E4–ORF3 oligomerization.

The results from coexpression of WT E4–ORF3 and the L103A mutant in *E. coli* indicate that L103A is able to limit the ability of the WT E4–ORF3 to multimerize. When the His_6-tagged L103A and untagged WT proteins were coexpressed in *E. coli*, the L103A protein clearly reduced the multimeric state of the WT protein. When expressed individually, the WT and L103A proteins had distinct gel filtration profiles that did not overlap to any significant extent. L103A also interfered with the ability of WT E4–ORF3 to reorganize PML-NB *in vivo*. Some of the nuclear punctæ that form when WT E4–ORF3 and the L103A mutant were coexpressed *in vivo* colocalized with PML-NB, but notably, PML-NB integrity was not disrupted. This result suggests that the ability of E4–ORF3 to interact with one or more components of PML-NB may be uncoupled from the activity of E4–ORF3 to relocalize PML-NB proteins into nuclear tracks. Since L103A is not able to colocalize with PML-NB, we infer that WT E4–ORF3 in nuclear punctæ with L103A directs PML-NB colocalization and that higher-order E4–ORF3 multimerization may be required for PML-NB disruption.

While the isoelectric point of E4–ORF3 at pH 5.5 would suggest a highly charged, hydrophilic protein at physiological pH, the protein possesses a large amount of hydrophobic character. The presence of a run of five leucine residues that are spaced between 2 and 4 residues apart near the C-terminus of E4–ORF3 may contribute to its poor solubility. Mutation of these residues to alanine has severe phenotypic consequences on viral growth (reference [50] and unpublished results). The roughly even spacing of these residues suggests that they may play a role in generating a short amphipathic alpha helix. This could mediate the self-associative properties of E4–ORF3, as suggested by the analytical ultracentrifugation data for L103A whose mutation maps to this region. Based on spectroscopic evidence, we demonstrated that L103A contains essentially identical secondary and tertiary structure compared to WT E4–ORF3; L103A protein largely forms dimers in solution and is deficient in the self-associative properties of WT E4–ORF3. Furthermore, the secondary structure of WT E4–ORF3 is invariant to multimerization, suggesting that the mechanism of E4–ORF3 polymerization involves the association of well-folded proteins. These results provide a basis for understanding the self-associative properties of E4–ORF3 as they pertain to the activity of the protein.

Chapter 6

General Conclusions

As part of this thesis, we presented investigations of three multivalent proteins. First, we studied two multivalent lectins with the aim of designing a lectin-based "mix-and-match" combinatorial scaffold. Such a scaffold would enable us to profile the glycan geometry of the glycoproteins on the surface of the pathogen HIV, and to explore novel carbohydrate specificity to inhibit viral infection. Finally, we conducted a biophysical investigation of the adenoviral protein E4–ORF3, which assembles a multivalent web in the host cell to capture antiviral proteins or to target them for degradation.

In the case of the HIV-inhibiting lectin Cyanovirin-N (CVN), we demonstrated that the stability of the wildtype molecule is not optimal, and that rational approaches can both detect destabilizing interactions, and design variants with improved stability. We were not, however, successful in producing correctly folded individual domains of CVN. Instead, we focused on the wildtype protein in order the understand the energetics of the fold, and to engineer variants which are more resistant to denaturation. One significant finding of our work is the uncharacteristically slow unfolding of CVN when exposed to intermediate concentrations of chemical denaturants. We observed that CVN populates both kinetic and equilibrium unfolding intermediates. While the structural nature of the intermediates is at the present time unknown, we can imagine that they correspond to partially-folded molecular ensembles, which contains some native-like features. Since CVN contains no proline residues, we hypothesize that disulfide bond isomerization may be responsible for its slow unfolding. As an intriguing extension of our work one can compare the thermodynamic stability and unfolding kinetics of CVN and variants (either natural or designed) which lack disulfide bonds. The second revelatory consequence of our work is the description of non-two-state equilibrium folding of CVN, despite the fact that steady-state tryptophan fluorescence and individual circular dichroism wavelengths gave no indication of this phenomenon. Only through collecting spectral data as a function of denaturant were we able to reveal this behavior. Future studies of CVN folding, potentially including real-time NMR, may shed further light onto the denaturation pathway of this protein.

In theory, the tetravalency of the antiviral lectin MVL provides more opportunities for combinatorial screening, when compared to the divalent CVN. Individual domains of MVL have no disulfide bonds and are assembled in a commonly-occurring α/β topology, and are thus expected more likely to fold correctly. We demonstrated that despite their homology, the individual domains of MVL differ in their stability. While the carboxy-terminal domain adopts a compact globular structure with a CD spectrum identical to the wildtype protein, the amino-terminal domain is unfolded. Efforts to induce or stabilize folding of the amino-terminal domain were not productive. Using biophysical techniques we determined that the C-terminal domain assembles into a dimer at µM concentrations. Based on the symmetry of MVL, we identified two possible dimeric configurations for $\mathrm{MVL_C}$. In order to predict which of the two configurations is more energetically favorable, we carried out a computational investigation which suggested that the domain dimerizes through a mainly-nonpolar interface present in the wildtype molecule. Based on the known energetics of carbohydrate binding by MVL, we hypothesized that the C-terminal domain alone may possess anti-HIV activity. This was confirmed by exposing four laboratory-adapted HIV strains to the excised domain and the wildtype protein in a neutralization assay. We found that the antiviral potency of the two inhibitors was indistinguishable, suggesting that that multivalency is not strictly necessary for potent HIV inhibition by MVL. This is in contrast to other well-characterized multivalent lectins such as CVN and griffithsin, which lose a significant fraction of their potency after monomerization.

In the context of carbohydrate recognition by MVL, we used molecular dynamics simulations and energy decomposition approaches to provide a structural and energetic basis for high-mannose and chitin-derived oligosaccharides by MVL. Our investigation uncovered a mixed mode of $\mathrm{GlcNAc_4}$ binding to MVL, which provided a way to interpret conflicting experimental data. Our studies highlighted the importance of considering dynamic ensembles when calculating the energetics of protein–carbohydrate interactions. Future computational investigations of carbohydrate recognition by MVL may focus on complex carbohydrates, since those are plentiful in the surface of HIV and are potential targets of MVL.

Finally, in consideration of the adenoviral protein E4–ORF3, we applied spectroscopic and biophysical techniques to characterize the structure and oligomerization state of this unusual viral protein. In cells, E4–ORF3 assembles into heterogeneous aggregates which capture host proteins, and prior to our investigation it was not known whether the aggregates are composed of well-folded, amorphous, or amyloid-like material. We demonstrated that the nuclear "tracks" produced as a consequence of viral infection are composed of well-folded dimeric building blocks. Using spectroscopic techniques, we determined that the secondary and tertiary structure of the recombinantly-produced wildtype E4–ORF3 and the non-functional L103A mutant were identical, and contained significant α-helical structure. Our study was the first biophysical investigation of this unique viral protein. Our work, in concert with the subsequently-solved X-ray structure of a related nonfunctional E4–ORF3 variant, provides a structural interpretation of the macromolecular assembly formed by E4–ORF3. Despite this, many questions about this protein remain unanswered. The most pressing question deals with the mechanism by which E4–ORF3 captures unrelated cellular substrates. One possibility, which we briefly explore, is that E4–ORF3 recognizes the SUMO post-translational modification common to a number of its substrates. Future work will likely address the structural details of this mechanism.

Appendix A

Biophysical Studies Reveal Weak Interaction Between L103A Mutant of E4–ORF3 and human SUMO1

Abstract

E4–ORF3, the gene product of the adenoviral early region 4 open reading frame 3 is, at 116 amino acids, modest in size. Upon viral infection, however, this protein forms heterogeneous supramolecular filaments within the host nucleus. These filaments are thought to capture cellular response proteins to viral infection, effectively disabling their function. The mechanism by which E4–ORF3 recognizes its various binding partners is presently not know. Herein, we propose a mechanism of E4–ORF3 function which involves specific recognition of the small ubiquitin-like modifier post-translational modification by E4–ORF3.

A.1 Introduction

The adenoviral protein E4–ORF3 (UniProt ID: `P04489`) is multifunctional and is able to affect a number of cellular processes by binding numerous host protein complexes. A prominent attribute of E4–ORF3 is its ability to form heterogeneous nuclear filaments (also known as "tracks"). Filament formation occurs even in the absence of all other Ad viral proteins; thus, self-interaction between E4–ORF3 building blocks, in addition to its association with host proteins are sufficient for filament assembly. It is thought that E4–ORF3 track formation creates a polyvalent nuclear scaffold which is responsible for re-localizing nuclear multiprotein complexes such as PML and MRN, as well as cytoplasmic proteins. E4–ORF3 has been determined to re-localize nearly a dozen host proteins; the mechanism of how this happens is presently unknown.

We previously characterized recombinant E4–ORF3 and its nonfunctional variant Leu103Ala (L103A, see Chapter 5). The L103A mutation imparts two significant phenotypes: first, the mutant is unable to form the hallmark tracks *in vivo*; second, the mutant loses the ability to interact with its binding partners *in vitro*. The known cellular targets of E4–ORF3 (PML and Daxx, E1B-

55K, the transcription factors TIF1γ and TIF1α, and components of the MRN complex (Mre11, Rad50 and Nbs1)) are unrelated in their sequences and structure. Some of these proteins are known to be SUMOylated; this large post-translational modification may provide a handle by which E4–ORF3 is able to capture proteins of dissimilar structure. In support of this hypothesis, recombinantly-prepared E4–ORF3 binds human SUMO1 in a pull-down assay, while the L103A mutant does not (P. Hearing, personal communication). We previously hypothesized that E4–ORF3 functions as a multivalent scaffold, by presenting a number of binding interfaces in the nucleus. It is also possible that a novel emergent interface is formed as a consequence of E4–ORF3 filament formation. How E4–ORF3 recognizes SUMO1, and how the L103A mutation interferes with this recognition is presently not known.

In this work, building upon our initial biophysical characterization of E4–ORF3, we present improved E4–ORF3 expression constructs and results of new biophysical experiments which aim to elucidate the mechanism by which E4–ORF3 recognizes its binding partners, focusing on its putative interaction with human SUMO1. Finally, building upon the recently-solved crystal structure of an E4–ORF3 variant [140], we propose a mechanism by which the L103A mutation weakens E4–ORF3 oligomerization and abrogates SUMO1 binding, and suggest experimental to validate the proposed mechanism.

A.2 Materials & Methods

Plasmid Constructs and Purification of E4–ORF3. The pProEx-HTb plasmid (ampicillin resistance) containing the genetic sequence of the L103A mutant of Ad5 E4–ORF3 was used to produce the protein in *E. coli*. This construct expressed L103A as a C-terminal fusion to a polyhistidine affinity tag, separated by a TEV protease cleavage site. Constructs containing two and four additional glycine residues at the TEV protease recognition sequence were constructed by site-directed mutagenesis. Recombinant L103A variants were expressed using the BL21(DE3) *E. coli* strain. Cultures were induced with 0.75 mM IPTG at an OD_{600} = 0.6–0.8 after cooling to room temperature. Inductions were performed for 20 h at 25 °C and cell pellets were frozen prior to protein purification.

E4–ORF3 variants were purified using a low-salt chromatographic sequence to avoid precipitating the recombinant protein. Cells were disrupted by sonication (ten 50% duty cycles) in lysis buffer (10 mM Tris-HCl, 10 mM βME, 10 mM imidazole, pH 7.5). Lysates were clarified by centrifugation at 40 kg for 30 min. The soluble fraction was passed through a 0.22 µm syringe filter and applied onto a 5 mL HisTrap affinity column (GE Healthcare). Following a wash of 10 column volumes of lysis buffer, and protein was eluted with the same buffer supplemented with 250 mM imidazole.

Fractions containing recombinant protein were combined and dialyzed against TEV protease cleavage buffer (20 mM Tris-HCl, 1 mM βME, pH 7.5) overnight at 4 °C. TEV protease cleavage was initiated by adding 250 µg of recombinant TEV protease, and the reaction progress was monitored by MALDI-TOF. Following cleavage, the reaction was passed through a HisTrap column pre-equilibrated with cleavage buffer, and the flow-through was collected. Undigested protein, the N-terminal E4–ORF3 peptide, as well as the TEV protease contain a poly-histidine tag; thus, they associate with the column, while untagged E4–ORF3 protein passes through. Size exclusion chromatography was carried out on a Superdex 200 16/60 column (GE Healthcare) in Buffer A (20 mM Tris-HCl, 0.5 mM TCEP, pH 7.5). Final protein yield was approximately 5.6 mg L^{-1} of expression culture.

Uniformly ^{15}N-labelled L103A was produced in M9 minimal medium supplemented with ^{15}N-labelled ammonium chloride as the only source of nitrogen, using standard protocols [79]. The protein was purified as above.

Purification of human SUMO1. The pET28a plasmid (kanamycin resistance) containing a gene encoding human SUMO1 fused to a C-terminal His$_6$-tag was obtained from Dr. Patrick Hearing. Recombinant protein was expressed using BL21(DE3): cultures were induced with 1 mM IPTG at an OD_{600} = 0.6–0.8 after cooling to room temperature. Inductions were performed for 20 h at 25 °C and cell pellets were frozen prior to protein purification. Cells were resuspended in lysis buffer (20 mM Tris-HCl, 1 mM βME, 20 mM imidazole, 350 mM NaCl, pH 8.0) and lysed by sonication. The lysate was centrifuged, and the soluble fraction was applied onto a 5 mL HisTrap affinity column. The column was washed with lysis buffer, and eluted with buffer containing 400 mM imidazole. Fractions containing protein were pooled and dialyzed overnight at 4 °C against SUMO protease cleavage buffer (20 mM Tris-HCl, 1 mM βME, pH 7.5) in the presence of 250 µg Ulp1 protease. The digest was re-applied onto 5 mL HisTrap column and the flow-through was collected. Size exclusion chromatography was carried out using a Superdex 200 16/60 column (GE Healthcare) in Buffer A (20 mM Tris-HCl, 0.5 mM TCEP, pH 7.5). The protein eluted as a monodisperse symmetric peak at 88 mL and its identity was confirmed using MALDI-TOF.

Gel Filtration Binding Assay. We employed a Superdex 75 (10/300) analytical gel filtration column, pre-equilibrated in gel filtration buffer (20 mM Tris-HCl, 0.5 mM TCEP, 100 mM NaCl, pH 7.5). Sodium chloride was included at 100 mM to screen potential interactions of the solute with the column resin; this salt concentration has been previously determined to have a minimal effect on precipitation of E4–ORF3. Proteins were chromatographed at a flow rate of 0.5 mL min^{-1} and their absorbance was detected at 280 nm.

NMR Spectroscopy. HSQC spectra of L103A+4G were collected using a Bruker 700 MHz conventional-probe spectrometer using 300 µM protein concentration at 300 K. Samples containing both L103A+4G and human SUMO1 were collected at two-fold SUMO1 excess (550 µM); all concentrations are provided in monomer units.

A.3 Results & Discussion

The L103A mutation in Ad5 E4–ORF3 is completely deleterious to protein function, and abolishes track formation *in vivo* (Chapter 5). We have previously determined that this mutation does not alter the overall secondary structure of E4–ORF3, but affects the protein quaternary structure/self-association. While L103A behaves as an ideal dimer *in vitro* at low µM concentrations, recombinantly-prepared wildtype protein is heterogeneous in solution and populates larger oligomers. We hypothesize that the dimer is the building block of larger E4–ORF3 oligomers.

The pProEx-HTb construct used to express L103A E4–ORF3 in our previous study may be sub-optimal for biophysical studies, as it installs a 24-residue N-terminal purification tag. This tag contains a TEV cleavage site, but cannot not be processed by the TEV protease, most likely due to poor accessibility of the recognition sequence; the tag contributes almost 2.9 kDa, corresponding to 18% of the protein MW. While the affinity tag was definitively demonstrated to not affect protein function *in vivo* at the resolution available to immunofluorescence (Chapter 5), it contains numerous charged residues and may influence the energetics of intermolecular association in biophysical experiments or contribute unfavorably to success of crystallization trials.

Optimization of L103A E4–ORF3 construct for *in vitro* biophysical studies. We set out to optimize the sequence of the TEV protease cleavage site with the goal of increasing proteolytic efficiency. We designed constructs containing two and four additional glycine residues after the P1′ position of the protease recognition sequence `ENLYFQG` [88]. While addition of two glycine residues had no observable effect on proteolytic efficiency, adding four glycine residues yielded protein which was readily processed by the protease (Figure A-1). The tagless L103A+4G protein was verified to contain identical secondary structure to previously-characterized E4–ORF3 variants using circular dichroism spectroscopy and this construct was used for all subsequent biophysical assays.

L103A does not bind SUMO1. The affinity of *in vitro* intermolecular association is often in the µM range or tighter. For such complexes, the upper bound of the equilibrium dissociation constant can be estimated through a non-equilibrium gel filtration assay. By measuring the retention time and peak volume of each molecular component and those of the mixture, one can estimate the population change due to intermolecular association. In general, the technique underestimates association due to sample dilution in the column or due to fast dissociation kinetics. We observed no binding between recombinantly-prepared dimeric L103A+4G (20 µM) and human SUMO1 (40 µM) in an analytical gel filtration assay (Figure A-2A); monomer concentrations are given for both components.

As the result of the limitations of gel filtration chromatography, the aforementioned result does not definitely rule out binding of SUMO1 by L103A. The binding event may be weak, or may have fast dissociation kinetics. NMR spectroscopy, on the other hand, under favorable exchange conditions (slow or fast), can provide residue-specific information on the binding event, and is an equilibrium technique. We prepared ^{15}N-containing L103A+4G protein, and collected HSQC

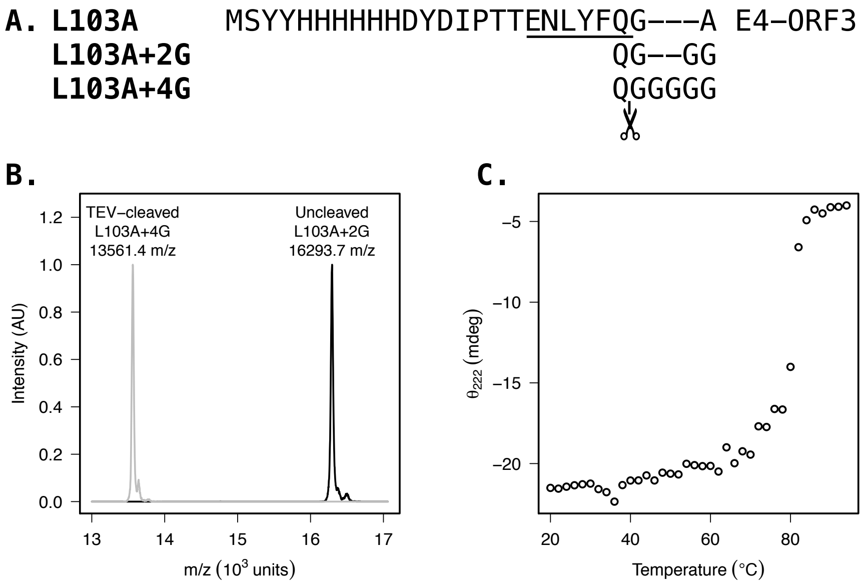


Figure A-1: **Optimizing the L103A construct for biophysical studies. A.** The amino acid sequence of the N-terminal purification tag used to produce L103A. The TEV protease recognition sequence is underlined, and the cleavage site is indicated. TEV protease cannot cleave the original construct; we designed and tested two constructs containing additional glycine residues C-terminal to the P1′ site. **B.** After a 24-hour incubation with TEV protease, complete cleavage of the L103A+4G protein was observed using MALDI-TOF (grey, expected 13542.6 m/z after cleavage), while the L103A+2G was not processed (black, expected 16265.5 before cleavage). **C.** The TEV-cleaved L103A+4G is stable to heat denaturation (monitored by CD spectroscopy), with an apparent denaturation midpoint above 80 °C; the protein monomer concentration was 27 µM. Apparent stability of E4–ORF3 is expected to be concentration-dependent

spectra of L103A+4G samples, and of L103A+4G in the presence of two-fold molar excess of human SUMO1 (Figure A-2B-C). The spectrum of L103A+4G was well-dispersed in both the proton and nitrogen dimensions. Due to peak crowding, we could not definitively determine whether the protein is a symmetric or asymmetric dimer. We observed that not all resonances were of the same linewidth and intensity, indicating that the molecule contains substructures of a different dynamical content relative to other parts of the molecule. Overlapping the HSQC spectra of L103A with and without hSUMO1 revealed no significant observable chemical shift changes. Thus, we conclude that L103A does not bind to SUMO1 with the ligand present in two-fold molar excess. These results suggest that the affinity of L103A for SUMO1 is greater than 1 mM, and are consistent with the inability of L103A to pull-down hSUMO1 or other known E4–ORF3 substrates in co-precipitation experiments.

E4–ORF3 may contain a SUMO interaction motif which is abolished by L103A mutation. The small ubiquitin-like modifier (SUMO) is a post-translational modification installed on lysine sidechain of various eukaryotic proteins [59, 157, 178]. Eukaryotes encode four SUMO

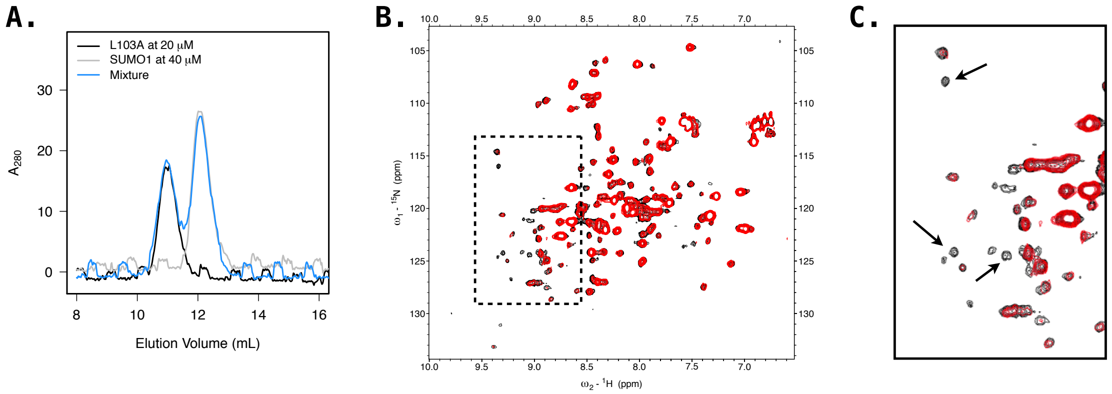


Figure A-2: **L103A binds human SUMO1 weakly, if at all. A.** An analytical gel filtration assay reveals no apparent binding between L103A and SUMO1. Samples were prepared at the stated concentrations, equilibrated at room temperature for several hours, and injected onto the column. **B.** An HSQC spectrum of *apo* L103A+4G (black) overlaid with the spectrum of L103A+4G in the presence of two-fold molar excess unlabelled SUMO1 (red). **C.** An expansion of the dashed HSQC region in panel **B**; arrows indicate resonance cross-peaks broadened by addition of the ligand. At the present time, it is not clear whether this interaction is specific.

homologs, which are typically grouped into three classes based on homology: SUMO1, SUMO2/3, and SUMO4. SUMO2/3 are 98% homologous and may have identical function; antibodies raised against one cross-react with the other homologue. SUMO1, on the other hand, is about 50% homologous to SUMO2/3, and is thought to have a distinct role. It is not known if SUMO4 is functional, or whether it is a *pseudo*-gene. Lysine sidechains can be poly-SUMOylated, analogous to ubiquitination; SUMO1 is thought to act as a poly-SUMO chain terminator.

Adenoviral infection is closely-linked with SUMOylation. For instance, the Ad protein E1B-55K is SUMOylated [47,93,104,190–192] and E4–ORF3 has been shown to regulate SUMOylation of the Mre11-Rad50-Nbs1 components [168]. The promyelocytic leukemia protein (PML) nuclear bodies, the localization of which is affected by Ad infection or E4–ORF3 transfection, are likely formed as a result of SUMO interactions with SUMO-interaction domains [162]. The associated factor Daxx is known to contain a SUMO-interaction motif (SIM) which mediates its association with PML [108]. SUMO-interaction motifs are typically characterized by a run of four nonpolar residues flanked by acidic or basic residues or serine residues which can be phosphorylated to control the SUMO–SIM interaction [59].

A recently-solved X-ray structure of an N82E E4–ORF3 variant, coupled with EM visualization, has revealed the mechanism by which E4–ORF3 assembles into nuclear polymers [140]. The nonfunctional N82E variant was reveled as a homodimer, in agreement with our previous observations (Chapter 5). The key finding which reveals the mechanism of E4–ORF3 polymerization involves the C-terminal tail (residues 99-116). This tail (Figure A-3A) is thought to mediate domain-swapping interactions which create a heterogeneous polymer network [140]. Tail residues which point toward the core of the homodimer are typically hydrophobic, and are highly conserved

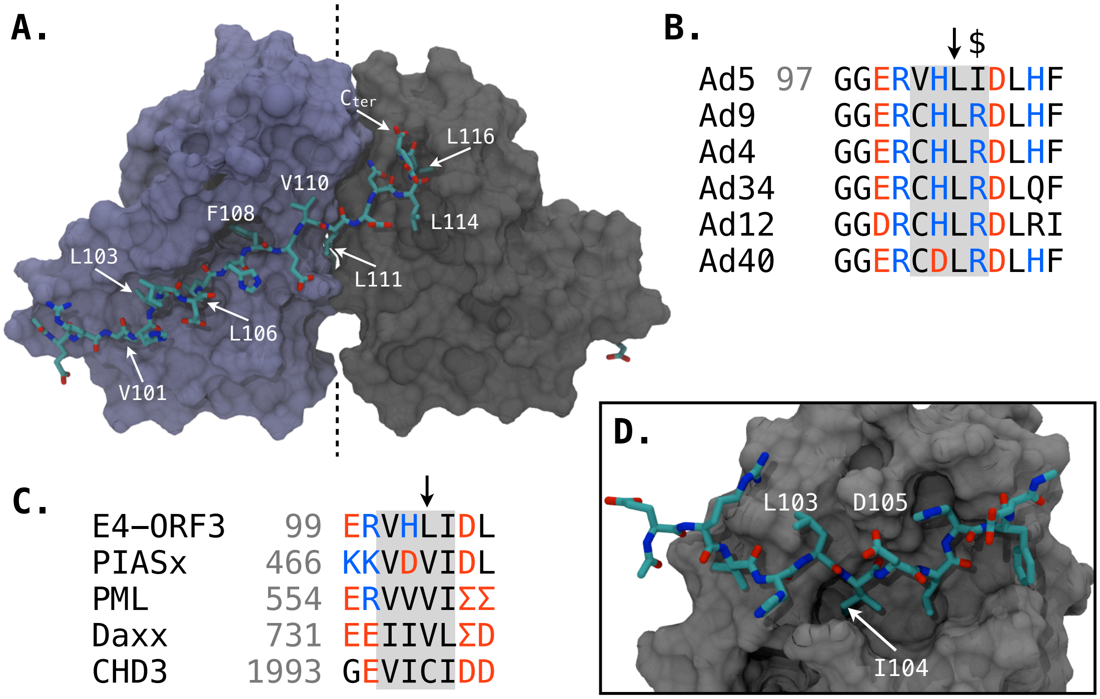


Figure A-3: **Proposed mechanism by E4–ORF3 function. A.** The crystal structure of the nonfunctional N82E variant of E4–ORF3(PDB entry 4DJB [140]). Consistent with the findings presented in Chapter 5, the structure reveals that E4–ORF3 homodimerizes, and that two long C-terminal tails (rendered in licorice, only one visible) occupy binding proclivities on the dimer surface. Tail residues V101, L103, L106, F108, V110, L111, and L116 point toward the core of the dimer; this tail is used by E4–ORF3 to polymerize the dimeric building blocks *via* a heterogeneous three-dimensional domain swapping mechanism. The L103A mutation likely weakens both intra- and intermolecular interactions involving this tail. **B.** The tail sequence is highly conserved across viral serotypes; L103, indicated with an arrow, is perfectly conserved. Intriguingly, adjacent H102 and I104 (identified with $) contain sequence variation. All adenoviral strains are able to form nuclear tracks. It has been shown that I104 is crucial for track function against the MRN complex, as the I104R mutation abolishes MRN re-localization by Ad5 E4–ORF3. **C.** The C-terminal tail of E4–ORF3 contains a sequence remarkably similar to known SUMO-interaction modifs of PIASx, PML, and Daxx (Σ represents phospho-serine). The SUMO-interaction motif is characterized by a stretch of hydrophobic amino acids flanked by neighboring basic or acidic groups. The site of the L103A mutation is indicated by an arrow. **D.** A homology model of human SUMO1 in complex the C-terminal tail of E4–ORF3. Model was constructed based on PDB entry 2ASQ [169] using MODELLER [149]. Remarkably, the putative E4–ORF3 SIM residue I104 is buried in a hydrophobic binding pocket on SUMO1. Mutation of this residue to Arg (as observed with Ad viral serotypes in panel **B**) will likely abolish SUMO1 binding. However, I104R should have no effect on track formation, due to the solvent accessibility of this residue.

(Figure A-3B); their mutation to alanine has severe consequences on nuclear track assembly [50].

Remarkably, the E4–ORF3 tail contains a string of residues which are exceptionally homologous to those of a known SUMO-interaction motif (Figure A-3C). We propose that this region of E4–ORF3 is responsible for binding the SUMO modification of numerous known cellular targets. E4–ORF3 proteins of numerous Adenoviral serotypes are all able to form nuclear tracks; however, they differ in their ability to interact with MRN [29, 168, 174]. A key reason for this appears to be the identity of the residue at position 104: clades containing Ile relocalize MRN, while those containing Arg at position 104 do not.

Additional experiments are required to validate or disprove the E4–ORF3 SIM hypothesis. However, this region has been shown to be the functional portion of the protein, and a number mutants which abrogate E4–ORF3 function map to residues 101 through 106. The most pressing question is whether an emergent interface is required to interact with cellular substrates, or whether the sequence of the tail alone is sufficient. Based on the structural model of E4–ORF3 tail in complex with SUMO1, we predict that L103A mutation weakens the interaction with SUMO1. This may work to explain the poor binding we observed between L103A variant of E4–ORF3 and human SUMO1.

Bibliography

[1] E. W. Adams, D. M. Ratner, H. R. Bokesch, J. B. McMahon, B. R. O'Keefe, and P. H. Seeberger. Oligosaccharide and glycoprotein microarrays as tools in HIV glycobiology; glycan-dependent gp120/protein interactions. *Chem Biol*, **11**(6):875–81, Jun 2004.

[2] K. B. Alexandre, E. S. Gray, B. E. Lambson, P. L. Moore, I. A. Choge, K. Mlisana, S. S. A. Karim, J. McMahon, B. O'Keefe, R. Chikwamba, and L. Morris. Mannose-rich glycosylation patterns on HIV-1 subtype C gp120 and sensitivity to the lectins griffithsin, cyanovirin-N and scytovirin. *Virology*, **402**(1):187–96, Jun 2010.

[3] B. Anil, B. Song, Y. Tang, and D. P. Raleigh. Exploiting the right side of the Ramachandran plot: substitution of glycines by D-alanine can significantly increase protein stability. *J Am Chem Soc*, **126**(41):13194–5, Oct 2004.

[4] F. D. Araujo, T. H. Stracker, C. T. Carson, D. V. Lee, and M. D. Weitzman. Adenovirus type 5 E4orf3 protein targets the Mre11 complex to cytoplasmic aggresomes. *J Virol*, **79**(17):11382–91, Sep 2005.

[5] A. Baker, K. J. Rohleder, L. A. Hanakahi, and G. Ketner. Adenovirus E4 34k and E1b 55k oncoproteins target host DNA ligase IV for proteasomal degradation. *J Virol*, **81**(13):7034–40, Jul 2007.

[6] J. Balzarini. Targeting the glycans of glycoproteins: a novel paradigm for antiviral therapy. *Nat Rev Microbiol*, **5**(8):583–97, Aug 2007.

[7] G. N. Barber. Innate immune DNA sensing pathways: STING, AIMII and the regulation of interferon production and inflammatory responses. *Curr Opin Immunol*, **23**(1):10–20, Feb 2011.

[8] J. P. Bardhan, M. D. Altman, D. J. Willis, S. M. Lippow, B. Tidor, and J. K. White. Numerical integration techniques for curved-element discretizations of molecule–solvent interfaces. *J Chem Phys*, **127**(1):014701, Jul 2007.

[9] L. G. Barrientos, F. Lasala, R. Delgado, A. Sanchez, and A. M. Gronenborn. Flipping the switch from monomeric to dimeric CV-N has little effect on antiviral activity. *Structure*, **12**(10):1799–1807, Oct 2004.

[10] L. G. Barrientos, J. M. Louis, I. Botos, T. Mori, Z. Han, B. R. O'Keefe, M. R. Boyd, A. Wlodawer, and A. M. Gronenborn. The domain-swapped dimer of cyanovirin-N is in a metastable folded state: reconciliation of X-ray and NMR structures. *Structure*, **10**(5):673–86, May 2002.

[11] L. G. Barrientos, J. M. Louis, J. Hung, T. H. Smith, B. R. O'Keefe, R. S. Gardella, T. Mori, M. R. Boyd, and A. M. Gronenborn. Design and initial characterization of a circular permuted variant of the potent HIV-inactivating protein cyanovirin-N. *Proteins*, **46**(2):153–60, Feb 2002.

[12] L. G. Barrientos, E. Matei, F. Lasala, R. Delgado, and A. M. Gronenborn. Dissecting carbohydrate–cyanovirin-N binding by structure-guided mutagenesis: functional implications for viral entry inhibition. *Protein Eng Des Sel*, **19**(12):525–35, Dec 2006.

[13] L. G. Barrientos, B. R. O'Keefe, M. Bray, A. Sanchez, A. M. Gronenborn, and M. R. Boyd. Cyanovirin-N binds to the viral surface glycoprotein, GP1,2 and inhibits infectivity of Ebola virus. *Antiviral Res*, **58**(1):47–56, Mar 2003.

[14] R. Bernardi and P. P. Pandolfi. Structure, dynamics and functions of promyelocytic leukaemia nuclear bodies. *Nat Rev Mol Cell Biol*, **8**(12):1006–16, Dec 2007.

[15] C. A. Bewley. Solution structure of a cyanovirin-N:Manα1-2Manα complex: structural basis for high-affinity carbohydrate-mediated binding to gp120. *Structure*, **9**(10):931–40, Oct 2001.

[16] C. A. Bewley, M. Cai, S. Ray, R. Ghirlando, M. Yamaguchi, and K. Muramoto. New carbohydrate specificity and HIV-1 fusion blocking activity of the cyanobacterial protein MVL: NMR, ITC and sedimentation equilibrium studies. *J Mol Biol*, **339**(4):901–14, Jun 2004.

[17] C. A. Bewley, K. R. Gustafson, M. R. Boyd, D. G. Covell, A. Bax, G. M. Clore, and A. M. Gronenborn. Solution structure of cyanovirin-N, a potent HIV-inactivating protein. *Nat Struct Biol*, **5**(7):571–8, Jul 1998.

[18] C. A. Bewley, S. Kiyonaka, and I. Hamachi. Site-specific discrimination by cyanovirin-N for α-linked trisaccharides comprising the three arms of Man$_8$ and Man$_9$. *J Mol Biol*, **322**(4):881–9, Sep 2002.

[19] C. A. Bewley and S. Otero-Quintero. The potent anti-HIV protein cyanovirin-N contains two novel carbohydrate binding sites that selectively bind to Man$_8$ D1D3 and Man$_9$ with nanomolar affinity: implications for binding to the HIV envelope protein gp120. *J Am Chem Soc*, **123**(17):3892–902, May 2001.

[20] M. Blaber, J. D. Lindstrom, N. Gassner, J. Xu, D. W. Heinz, and B. W. Matthews. Energetic cost and structural consequences of burying a hydroxyl group within the core of a protein determined from Ala→Ser and Val→Thr substitutions in T4 lysozyme. *Biochemistry*, **32**(42):11363–73, Oct 1993.

[21] P. Blanchette, K. Kindsmüller, P. Groitl, F. Dallaire, T. Speiseder, P. E. Branton, and T. Dobner. Control of mRNA export by adenovirus E4orf6 and E1B55K proteins during productive infection requires E4orf6 ubiquitin ligase activity. *J Virol*, **82**(6):2642–51, Mar 2008.

[22] O. Blixt, S. Head, T. Mondala, C. Scanlan, M. E. Huflejt, R. Alvarez, M. C. Bryan, F. Fazio, D. Calarese, J. Stevens, N. Razi, D. J. Stevens, J. J. Skehel, I. van Die, D. R. Burton, I. A. Wilson, R. Cummings, N. Bovin, C.-H. Wong, and J. C. Paulson. Printed covalent glycan array for ligand profiling of diverse glycan binding proteins. *Proc Natl Acad Sci USA*, **101**(49):17033–8, Dec 2004.

[23] S. Boresch, G. Archontis, and M. Karplus. Free energy simulations: the meaning of the individual contributions from a component analysis. *Proteins*, **20**(1):25–33, Sep 1994.

[24] S. Boresch and M. Karplus. The meaning of component analysis: decomposition of the free energy in terms of specific interactions. *J Mol Biol*, **254**(5):801–7, Dec 1995.

[25] J. U. Bowie, J. F. Reidhaar-Olson, W. A. Lim, and R. T. Sauer. Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science*, **247**(4948):1306–10, Mar 1990.

[26] M. R. Boyd, K. R. Gustafson, J. B. McMahon, R. H. Shoemaker, B. R. O'Keefe, T. Mori, R. J. Gulakowski, L. Wu, M. I. Rivera, C. M. Laurencot, M. J. Currens, J. H. Cardellina, R. W. Buckheit, P. L. Nara, L. K. Pannell, R. C. Sowder, and L. E. Henderson. Discovery of cyanovirin-N, a novel human immunodeficiency virus-inactivating protein that binds viral surface envelope glycoprotein gp120: potential applications to microbicide development. *Antimicrob Agents Chemother*, **41**(7):1521–30, Jul 1997.

[27] B. R. Brooks, C. L. Brooks III, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. CHARMM: The Biomolecular Simulation Program. *J Comput Chem*, **30**(10, Sp. Iss. SI):1545–614, May 2009.

[28] N. Carrascal and D. F. Green. Energetic decomposition with the generalized-born and Poisson–Boltzmann solvent models: lessons from association of G-protein components. *J Phys Chem B*, **114**(15):5096–116, Apr 2010.

[29] C. T. Carson, N. I. Orazio, D. V. Lee, J. Suh, S. Bekker-Jensen, F. D. Araujo, S. S. Lakdawala, C. E. Lilley, J. Bartek, J. Lukas, and M. D. Weitzman. Mislocalization of the MRN complex prevents ATR signaling during adenovirus infection. *EMBO J*, **28**(6):652–62, Mar 2009.

[30] T. Carvalho, J. S. Seeler, K. Ohman, P. Jordan, U. Pettersson, G. Akusjärvi, M. Carmo-Fonseca, and A. Dejean. Targeting of adenovirus E1A and E4-ORF3 proteins to nuclear matrix-associated PML bodies. *J Cell Biol*, **131**(1):45–56, Oct 1995.

[31] L. C. Chang and C. A. Bewley. Potent inhibition of HIV-1 fusion by cyanovirin-N requires only a single high affinity carbohydrate binding site: characterization of low affinity carbohydrate binding site knockout mutants. *J Mol Biol*, **318**(1):1–8, Apr 2002.

[32] T. K. Chaudhuri, M. Arai, T. P. Terada, T. Ikura, and K. Kuwajima. Equilibrium and kinetic studies on folding of the authentic and recombinant forms of human α-lactalbumin by circular dichroism spectroscopy. *Biochemistry*, **39**(50):15643–51, Dec 2000.

[33] T. K. Chaudhuri, K. Horii, T. Yoda, M. Arai, S. Nagata, T. P. Terada, H. Uchiyama, T. Ikura, K. Tsumoto, H. Kataoka, M. Matsushima, K. Kuwajima, and I. Kumagai. Effect of the extra N-terminal methionine residue on the stability and folding of recombinant α-lactalbumin expressed in Escherichia coli. *J Mol Biol*, **285**(3):1179–94, Jan 1999.

[34] J.-H. Cho, S. Sato, and D. P. Raleigh. Thermodynamics and kinetics of non-native interactions in protein folding: a single point mutant significantly stabilizes the N-terminal domain of L9 by modulating non-native interactions in the denatured state. *J Mol Biol*, **338**(4):827–37, May 2004.

[35] D. M. Colleluori, D. Tien, F. Kang, T. Pagliei, R. Kuss, T. McCormick, K. Watson, K. McFadden, I. Chaiken, R. W. Buckheit, and J. W. Romano. Expression, purification, and characterization of recombinant cyanovirin-N for vaginal anti-HIV microbicide development. *Protein Expr Purif*, **39**(2):229–36, Feb 2005.

[36] W. Cornell, P. Cieplak, C. Bayly, I. Gould, K. Merz, D. Ferguson, D. Spellmeyer, T. Fox, J. Caldwell, and P. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc*, **118**(9):5179–97, Jan 1996.

[37] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner. WebLogo: a sequence logo generator. *Genome Res*, **14**(6):1188–90, Jun 2004.

[38] B. I. Dahiyat and S. L. Mayo. Probing the role of packing specificity in protein design. *Proc Natl Acad Sci USA*, **94**(19):10172–7, Sep 1997.

[39] F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer, and A. Bax. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR*, **6**(3):277–93, Nov 1995.

[40] J. Desmet, M. Maeyer, B. Hazes, and I. Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, **356**:539–42, April 1992.

[41] B. Dey, D. L. Lerner, P. Lusso, M. R. Boyd, J. H. Elder, and E. A. Berger. Multiple antiviral activities of cyanovirin-N: blocking of human immunodeficiency virus type 1 gp120 interaction with CD4 and coreceptor and inhibition of diverse enveloped viruses. *J Virology*, **74**(10):4562–9, May 2000.

[42] K. A. Dill and D. Shortle. Denatured states of proteins. *Annu Rev Biochem*, **60**:795–825, Jan 1991.

[43] V. Doucas, A. M. Ishov, A. Romo, H. Juguilon, M. D. Weitzman, R. M. Evans, and G. G. Maul. Adenovirus replication is coupled with the dynamic properties of the PML nuclear structure. *Genes Dev*, **10**(2):196–207, Jan 1996.

[44] R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu, and D. E. Shaw. Biomolecular simulation: a computational microscope for molecular biology. *Annu Rev Biophys*, **41**:429–52, Jun 2012.

[45] D. M. Eckert and P. S. Kim. Mechanisms of viral membrane fusion and its inhibition. *Annu Rev Biochem*, **70**:777–810, Jan 2001.

[46] D. Eisenberg and A. D. McLachlan. Solvation energy in protein folding and binding. *Nature*, **319**(6050):199–203, Jan 1986.

[47] C. Endter, J. Kzhyshkowska, R. Stauber, and T. Dobner. SUMO-1 modification required for transformation by adenovirus type 5 early region 1B 55-kda oncoprotein. *Proc Natl Acad Sci USA*, **98**(20):11312–7, Sep 2001.

[48] M. T. Esser, T. Mori, I. Mondor, Q. J. Sattentau, B. Dey, E. A. Berger, M. R. Boyd, and J. D. Lifson. Cyanovirin-N binds to gp120 to interfere with CD4-dependent human immunodeficiency virus type 1 virion binding, fusion, and infectivity but does not affect the CD4 binding site on gp120 or soluble CD4-induced conformational changes in gp120. *J Virol*, **73**(5):4360–71, May 1999.

[49] J. D. Evans and P. Hearing. Distinct roles of the adenovirus E4 ORF3 protein in viral DNA replication and inhibition of genome concatenation. *J Virol*, **77**(9):5295–304, May 2003.

[50] J. D. Evans and P. Hearing. Relocalization of the Mre11-Rad50-Nbs1 complex by the adenovirus E4 ORF3 protein is required for viral replication. *J Virol*, **79**(10):6207–15, May 2005.

[51] R. D. Everett and M. K. Chelbi-Alix. PML and PML nuclear bodies: implications in antiviral defense. *Biochimie*, **89**(6-7):819–30, Jan 2007.

[52] C. Fasting, C. A. Schalley, M. Weber, O. Seitz, S. Hecht, B. Koksch, J. Dernedde, C. Graf, E.-W. Knapp, and R. Haag. Multivalency as a chemical organization and action principle. *Angew Chem Int Ed Engl*, **51**(42):10472–98, Oct 2012.

[53] T. Feizi, F. Fazio, W. Chai, and C. H. Wong. Carbohydrate microarrays—a new set of technologies at the frontiers of glycomics. *Curr Opin Struc Biol*, **13**(5):637–45, Oct 2003.

[54] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, and A. Bateman. The Pfam protein families database. *Nucleic Acids Res*, **38**(Database issue):D211–22, Jan 2010.

[55] N. A. Forrester, R. N. Patel, T. Speiseder, P. Groitl, G. G. Sedgwick, N. J. Shimwell, R. I. Seed, P. Ó. Catnaigh, C. J. McCabe, G. S. Stewart, T. Dobner, R. J. A. Grand, A. Martin, and A. S. Turnell. Adenovirus E4orf3 targets transcriptional intermediary factor 1γ for proteasome-dependent degradation during infection. *J Virol*, **86**(6):3167–79, Mar 2012.

[56] R. Fromme, Z. Katiliene, B. Giomarelli, F. Bogani, J. M. Mahon, T. Mori, P. Fromme, and G. Ghirlanda. A monovalent mutant of cyanovirin-N provides insight into the role of multiple interactions with gp120 for antiviral activity. *Biochemistry*, **46**(32):9199–207, Aug 2007.

[57] Y. K. Fujimoto and D. F. Green. Carbohydrate recognition by the antiviral lectin cyanovirin-N. *J Am Chem Soc*, **134**(48):19639–51, Dec 2012.

[58] Y. K. Fujimoto, R. N. Terbush, V. Patsalo, and D. F. Green. Computational models explain the oligosaccharide specificity of cyanovirin-N. *Protein Sci*, **17**(11):2008–14, Nov 2008.

[59] J. R. Gareau and C. D. Lima. The SUMO pathway: emerging mechanisms that shape specificity, conjugation and recognition. *Nat Rev Mol Cell Biol*, **11**(12):861–71, Dec 2010.

[60] M.-C. Geoffroy and M. K. Chelbi-Alix. Role of promyelocytic leukemia protein in host antiviral defense. *J Interferon Cytokine Res*, **31**(1):145–58, Jan 2011.

[61] M. K. Gilson, A. Rashin, R. Fine, and B. Honig. On the calculation of electrostatic interactions in proteins. *J Mol Biol*, **184**(3):503–16, Aug 1985.

[62] R. Gonzalez, W. Huang, R. Finnen, C. Bragg, and S. J. Flint. Adenovirus E1B 55-kilodalton protein is required for both regulation of mRNA export and efficient entry into the late phase of infection in normal human fibroblasts. *J Virol*, **80**(2):964–74, Jan 2006.

[63] F. D. Goodrum and D. A. Ornelles. Roles for the E4 orf6, orf3, and E1B 55-kilodalton proteins in cell cycle-independent adenovirus replication. *J Virol*, **73**(9):7474–88, Sep 1999.

[64] D. B. Gordon and S. L. Mayo. Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J Comput Chem*, **19**(13):1505–14, 1998.

[65] D. F. Green. Optimized parameters for continuum solvation calculations with carbohydrates. *J Phys Chem B*, **112**(16):5238–49, Apr 2008.

[66] D. F. Green, A. T. Dennis, P. S. Fam, B. Tidor, and A. Jasanoff. Rational design of new binding specificity by simultaneous mutagenesis of calmodulin and a target peptide. *Biochemistry*, **45**(41):12547–59, Oct 2006.

[67] D. F. Green and B. Tidor. Evaluation of electrostatic interactions. *Curr Protocols in Bioinformatics*, 2002.

[68] N. J. Greenfield. Determination of the folding of proteins as a function of denaturants, osmolytes or ligands using circular dichroism. *Nat Prot*, **1**(6):2733–41, Jan 2006.

[69] N. J. Greenfield. Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions. *Nat Prot*, **1**(6):2527–35, Jan 2006.

[70] K. R. Gustafson, R. C. Sowder, L. E. Henderson, J. H. Cardellina, J. B. McMahon, U. Rajamani, L. K. Pannell, and M. R. Boyd. Isolation, primary sequence determination, and disulfide bond structure of cyanovirin-N, an anti-HIV (human immunodeficiency virus) protein from the cyanobacterium *Nostoc ellipsosporum*. *Biochem Biophys Res Commun*, **238**(1):223–8, Sep 1997.

[71] Z. Han, C. Xiong, T. Mori, and M. R. Boyd. Discovery of a stable dimeric mutant of cyanovirin-N (CV-N) from a T7 phage-displayed CV-N mutant library. *Biochem Biophys Res Commun*, **292**(4):1036–1043, Apr 2002.

[72] J. N. Harada, A. Shevchenko, A. Shevchenko, D. C. Pallas, and A. J. Berk. Analysis of the adenovirus E1B-55K-anchored proteome reveals its link to ubiquitination machinery. *J Virol*, **76**(18):9194–206, Sep 2002.

[73] P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber, and P. S. Kim. High-resolution protein design with backbone freedom. *Science*, **282**(5393):1462–7, Nov 1998.

[74] B. Hazes and B. W. Dijkstra. Model building of disulfide bonds in proteins with known three-dimensional structure. *Protein Eng*, **2**(2):119–25, Jul 1988.

[75] Z. S. Hendsch, T. Jonsson, R. T. Sauer, and B. Tidor. Protein stabilization by removal of unsatisfied polar groups: computational approaches and experimental tests. *Biochemistry*, **35**(24):7621–5, Jun 1996.

[76] Z. S. Hendsch and B. Tidor. Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Sci*, **3**(2):211–26, Feb 1994.

[77] Z. S. Hendsch and B. Tidor. Electrostatic interactions in the GCN4 leucine zipper: substantial contributions arise from intramolecular interactions enhanced on binding. *Protein Sci*, **8**(7):1381–92, Jul 1999.

[78] E. Henry and J. Hofrichter. Singular value decomposition – application to analysis of experimental data. *Meth Enzymol*, **210**:129–92, Jan 1992.

[79] J. M. Hill. NMR screening for rapid protein characterization in structural proteomics. *Methods Mol Biol*, **426**:437–46, Jan 2008.

[80] B. Honig and A. Nicholls. Classical electrostatics in biology and chemistry. *Science*, **268**(5214):1144–9, May 1995.

[81] K.-L. Hsu and L. K. Mahal. Sweet tasting chips: microarray-based analysis of glycans. *Curr Opin Chem Biol*, **13**(4):427–32, Oct 2009.

[82] K.-L. Hsu, K. T. Pilobello, and L. K. Mahal. Analyzing the dynamic bacterial glycome with a lectin microarray approach. *Nat Chem Biol*, **2**(3):153–7, Mar 2006.

[83] S. J. Hubbard and J. M. Thornton. NACCESS, 1993.

[84] W. Humphrey, A. Dalke, and K. Schulten. VMD: visual molecular dynamics. *J Mol Graph*, **14**(1):33–8, Feb 1996.

[85] C. A. Hunter and H. L. Anderson. What is cooperativity? *Angew Chem Int Ed Engl*, **48**(41):7488–99, Jan 2009.

[86] E. G. Hutchinson and J. M. Thornton. A revised set of potentials for beta-turn formation in proteins. *Protein Sci*, **3**(12):2207–16, Dec 1994.

[87] A. Imberty, E. Mitchell, and M. Wimmerova. Structural basis of high-affinity glycan recognition by bacterial and fungal lectins. *Curr Opin Struc Biol*, **15**(5):525–34, Oct 2005.

[88] R. B. Kapust, J. Tözsér, T. D. Copeland, and D. S. Waugh. The P1′ specificity of tobacco etch virus protease. *Biochem Biophys Res Commun*, **294**(5):949–55, Jun 2002.

[89] K. A. Karen, P. J. Hoey, C. S. H. Young, and P. Hearing. Temporal regulation of the Mre11-Rad50-Nbs1 complex during adenovirus infection. *J Virol*, **83**(9):4565–73, May 2009.

[90] J. R. Keeffe, P. N. P. Gnanapragasam, S. K. Gillespie, J. Yong, P. J. Bjorkman, and S. L. Mayo. Designed oligomers of cyanovirin-N show enhanced HIV neutralization. *Proc Natl Acad Sci USA*, **108**(34):14079–84, Aug 2011.

[91] J.-C. Kehr, Y. Zilliges, A. Springer, M. D. Disney, D. D. Ratner, C. Bouchier, P. H. Seeberger, N. T. de Marsac, and E. Dittmann. A mannan binding lectin is involved in cell–cell attachment in a toxic strain of *Microcystis aeruginosa*. *Mol Microbiol*, **59**(3):893–906, Feb 2006.

[92] B. S. Kelley, L. C. Chang, and C. A. Bewley. Engineering an obligate domain-swapped dimer of cyanovirin-N with enhanced anti-HIV activity. *J Am Chem Soc*, **124**(13):3210–1, Apr 2002.

[93] K. Kindsmüller, P. Groitl, B. Härtl, P. Blanchette, J. Hauber, and T. Dobner. Intranuclear targeting and nuclear export of the adenovirus E1B-55K protein are regulated by SUMO1 conjugation. *Proc Natl Acad Sci USA*, **104**(16):6684–9, Apr 2007.

[94] L. M. I. Koharudin, A. R. Viscomi, J.-G. Jee, S. Ottonello, and A. M. Gronenborn. The evolutionarily conserved family of cyanovirin-N homologs: structures and carbohydrate specificity. *Structure*, **16**(4):570–84, Apr 2008.

[95] L. M. I. Koharudin, A. R. Viscomi, B. Montanini, M. J. Kershaw, N. J. Talbot, S. Ottonello, and A. M. Gronenborn. Structure-function analysis of a CVNH-LysM lectin expressed during plant infection by the rice blast fungus *Magnaporthe oryzae*. *Structure*, **19**(5):662–74, May 2011.

[96] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, and T. E. Cheatham. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res*, **33**(12):889–97, Dec 2000.

[97] M. Kuttel, J. W. Brady, and K. J. Naidoo. Carbohydrate solution simulations: producing a force field with experimentally consistent primary alcohol rotational frequencies and populations. *J Comput Chem*, **23**(13):1236–43, Oct 2002.

[98] S. S. Lakdawala, R. A. Schwartz, K. Ferenchak, C. T. Carson, B. P. McSharry, G. W. Wilkinson, and M. D. Weitzman. Differential requirements of the C-terminus of Nbs1 in suppressing adenovirus DNA replication and promoting concatemer formation. *J Virol*, **82**(17):8362–72, Sep 2008.

[99] I. Lasters, M. D. Maeyer, and J. Desmet. Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein side chains. *Protein Eng*, **8**(8):815–22, Aug 1995.

[100] R. F. Latypov, H. Cheng, N. A. Roder, J. Zhang, and H. Roder. Structural characterization of an equilibrium unfolding intermediate in cytochrome *c*. *J Mol Biol*, **357**(3):1009–25, Mar 2006.

[101] D. V. Laurents and R. L. Baldwin. Characterization of the unfolding pathway of hen egg white lysozyme. *Biochemistry*, **36**(6):1496–504, Feb 1997.

[102] A. R. Leach and A. P. Lemon. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins*, **33**(2):227–39, Nov 1998.

[103] K. N. Leppard and R. D. Everett. The adenovirus type 5 E1b 55K and E4 Orf3 proteins associate in infected cells and affect ND10 components. *J Gen Virol*, **80**(Pt 4):997–1008, Apr 1999.

[104] K. J. Lethbridge, G. E. Scott, and K. N. Leppard. Nuclear matrix localization and SUMO-1 modification of adenovirus type 5 E1b 55K protein are controlled by E4 Orf6 protein. *J Gen Virol*, **84**(Pt 2):259–68, Feb 2003.

[105] R. M. Levy, L. Y. Zhang, E. Gallicchio, and A. K. Felts. On the nonpolar hydration free energy of proteins: surface area and continuum solvent models for the solute–solvent interaction energy. *J Am Chem Soc*, **125**(31):9523–30, Aug 2003.

[106] P. Li, S. Banjade, H.-C. Cheng, S. Kim, B. Chen, L. Guo, M. Llaguno, J. V. Hollingsworth, D. S. King, S. F. Banani, P. S. Russo, Q.-X. Jiang, B. T. Nixon, and M. K. Rosen. Phase transitions in the assembly of multivalent signalling proteins. *Nature*, **483**(7389):336–40, Mar 2012.

[107] D. Liao, A. Yu, and A. M. Weiner. Coexpression of the adenovirus 12 E1B 55 kDa oncoprotein and cellular tumor suppressor p53 is sufficient to induce metaphase fragility of the human RNU2 locus. *Virology*, **254**(1):11–23, Feb 1999.

[108] D.-Y. Lin, Y.-S. Huang, J.-C. Jeng, H.-Y. Kuo, C.-C. Chang, T.-T. Chao, C.-C. Ho, Y.-C. Chen, T.-P. Lin, H.-I. Fang, C.-C. Hung, C.-S. Suen, M.-J. Hwang, K.-S. Chang, G. G. Maul, and H.-M. Shih. Role of SUMO-interacting motif in Daxx SUMO modification, subnuclear localization, and repression of SUMOylated transcription factors. *Mol Cell*, **24**(3):341–54, Nov 2006.

[109] S. M. Lippow, K. D. Wittrup, and B. Tidor. Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat Biotechnol*, **25**(10):1171–6, Oct 2007.

[110] H. Lis and N. Sharon. Lectins as molecules and as tools. *Annu Rev Biochem*, **55**:35–67, Jan 1986.

[111] Y. Liu, A. Shevchenko, A. Shevchenko, and A. J. Berk. Adenovirus exploits the cellular aggresome response to accelerate inactivation of the MRN complex. *J Virol*, **79**(22):14004–16, Nov 2005.

[112] S. W. Lockless and R. Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**(5438):295–9, Oct 1999.

[113] L. L. Looger and H. W. Hellinga. Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. *J Mol Biol*, **307**(1):429–45, Mar 2001.

[114] S. C. Lovell, J. M. Word, J. S. Richardson, and D. C. Richardson. The penultimate rotamer library. *Proteins*, **40**(3):389–408, Aug 2000.

[115] A. MacKerel Jr., C. Brooks III, L. Nilsson, B. Roux, Y. Won, and M. Karplus. *CHARMM: The energy function and its parameterization with an overview of the program*, volume 1 of *The Encyclopedia of Computational Chemistry*, pages 271–7. John Wiley & Sons, 1998.

[116] A. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. Evanseck, M. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. Lau, C. Mattos, S. Michnick, T. Ngo, D. Nguyen, B. Prodhom, W. Reiher, B. Roux, M. Schlenkrich, J. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B*, **102**(18):3586–616, 1998.

[117] M. D. Maeyer, J. Desmet, and I. Lasters. The dead-end elimination theorem: mathematical aspects, implementation, optimizations, evaluation, and performance. *Methods Mol Biol*, **143**:265–304, Jan 2000.

[118] S. M. Malakauskas and S. L. Mayo. Design, structure and stability of a hyperthermophilic protein variant. *Nat Struct Biol*, **5**(6):470–5, Jun 1998.

[119] M. Mammen, S.-K. Choi, and G. M. Whitesides. Polyvalent interactions in biological systems: Implications for design and use of multivalent ligands and inhibitors. *Angewandte Chemie International Edition*, **37**(20):2754–94, Nov 1998.

[120] D. J. Mandell and T. Kortemme. Computer-aided design of functional protein interactions. *Nat Chem Biol*, **5**(11):797–807, Nov 2009.

[121] E. Matei, W. Furey, and A. M. Gronenborn. Solution and crystal structures of a sugar binding site mutant of cyanovirin-N: no evidence of domain swapping. *Structure*, **16**(8):1183–94, Aug 2008.

[122] E. Matei, A. Zheng, W. Furey, J. Rose, C. Aiken, and A. M. Gronenborn. Anti-HIV activity of defective cyanovirin-N mutants is restored by dimerization. *J Biol Chem*, **285**(17):13057–65, Apr 2010.

[123] S. S. Mathew and E. Bridge. The cellular Mre11 protein interferes with adenovirus E4 mutant DNA replication. *Virology*, **365**(2):346–55, Sep 2007.

[124] S. S. Mathew and E. Bridge. Nbs1-dependent binding of Mre11 to adenovirus E4 mutant viral DNA is important for inhibiting DNA replication. *Virology*, **374**(1):11–22, Apr 2008.

[125] Y. K. Mok, E. L. Elisseeva, A. R. Davidson, and J. D. Forman-Kay. Dramatic stabilization of an SH3 domain by a single substitution: roles of the folded and unfolded states. *J Mol Biol*, **307**(3):913–28, Mar 2001.

[126] T. Mori, L. G. Barrientos, Z. Han, A. M. Gronenborn, J. A. Turpin, and M. R. Boyd. Functional homologs of cyanovirin-N amenable to mass production in prokaryotic and eukaryotic hosts. *Protein Expr Purif*, **26**(1):42–9, Oct 2002.

[127] T. Mori, K. R. Gustafson, L. K. Pannell, R. H. Shoemaker, L. Wu, J. B. McMahon, and M. R. Boyd. Recombinant production of cyanovirin-N, a potent human immunodeficiency virus-inactivating protein derived from a cultured cyanobacterium. *Protein Expr Purif*, **12**(2):151–8, Mar 1998.

[128] T. Mori, R. H. Shoemaker, R. J. Gulakowski, B. L. Krepps, J. B. McMahon, K. R. Gustafson, L. K. Pannell, and M. R. Boyd. Analysis of sequence requirements for biological activity of cyanovirin-N, a potent HIV (human immunodeficiency virus)-inactivating protein. *Biochem Biophys Res Commun*, **238**(1):218–22, Sep 1997.

[129] T. Moulaei, S. R. Shenoy, B. Giomarelli, C. Thomas, J. B. McMahon, Z. Dauter, B. R. O'Keefe, and A. Wlodawer. Monomerization of viral entry inhibitor griffithsin elucidates the relationship between multivalent binding to carbohydrates and anti-HIV activity. *Structure*, **18**(9):1104–15, Sep 2010.

[130] M. Munson, R. O'Brien, J. M. Sturtevant, and L. Regan. Redesigning the hydrophobic core of a four-helix-bundle protein. *Protein Sci*, **3**(11):2015–22, Nov 1994.

[131] J. K. Myers, C. N. Pace, and J. M. Scholtz. Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci*, **4**(10):2138–48, Oct 1995.

[132] S. Nauli, B. Kuhlman, and D. Baker. Computer-based redesign of a protein folding pathway. *Nat Struct Biol*, **8**(7):602–5, Jul 2001.

[133] E. Neher. How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci USA*, **91**(1):98–102, Jan 1994.

[134] M. Nevels, B. Täuber, E. Kremmer, T. Spruss, H. Wolf, and T. Dobner. Transforming potential of the adenovirus type 5 E4orf3 protein. *J Virol*, **73**(2):1591–600, Feb 1999.

[135] M. Nina, D. Beglov, and B. Roux. Atomic radii for continuum electrostatics calculations based on molecular dynamics free energy simulations. *The Journal of Physical Chemistry B*, **101**(26):5239–48, 1997.

[136] K. Nordqvist, K. Ohman, and G. Akusjärvi. Human adenovirus encodes two proteins which have opposite effects on accumulation of alternatively spliced mRNAs. *Mol Cell Biol*, **14**(1):437–45, Jan 1994.

[137] K. Ohman, K. Nordqvist, and G. Akusjärvi. Two adenovirus proteins with redundant activities in virus growth facilitates tripartite leader mRNA accumulation. *Virology*, **194**(1):50–8, May 1993.

[138] B. R. O'Keefe, D. F. Smee, J. A. Turpin, C. J. Saucedo, K. R. Gustafson, T. Mori, D. Blakeslee, R. Buckheit, and M. R. Boyd. Potent anti-influenza activity of cyanovirin-N and interactions with viral hemagglutinin. *Antimicrob Agents Chemother*, **47**(8):2518–25, Aug 2003.

[139] N. I. Orazio, C. M. Naeger, J. Karlseder, and M. D. Weitzman. The adenovirus E1b55K/ E4orf6 complex induces degradation of the Bloom helicase during infection. *J Virol*, **85**(4):1887–92, Feb 2011.

[140] H. D. Ou, W. Kwiatkowski, T. J. Deerinck, A. Noske, K. Y. Blain, H. S. Land, C. Soria, C. J. Powers, A. P. May, X. Shu, R. Y. Tsien, J. A. J. Fitzpatrick, J. A. Long, M. H. Ellisman, S. Choe, and C. C. O'Shea. A structural basis for the assembly and functions of a viral polymer that inactivates multiple tumor suppressors. *Cell*, **151**(2):304–19, Oct 2012.

[141] R. Pejchal, K. J. Doores, L. M. Walker, R. Khayat, P.-S. Huang, S.-K. Wang, R. L. Stanfield, J.-P. Julien, A. Ramos, M. Crispin, R. Depetris, U. Katpally, A. Marozsan, A. Cupo, S. Maloveste, Y. Liu, R. McBride, Y. Ito, R. W. Sanders, C. Ogohara, J. C. Paulson, T. Feizi, C. N. Scanlan, C.-H. Wong, J. P. Moore, W. C. Olson, A. B. Ward, P. Poignard, W. R. Schief, D. R. Burton, and I. A. Wilson. A potent and broad neutralizing antibody recognizes and penetrates the HIV glycan shield. *Science*, **334**(6059):1097–103, Nov 2011.

[142] R. Percudani, B. Montanini, and S. Ottonello. The anti-HIV cyanovirin-N domain is evolutionarily conserved and occurs as a protein module in eukaryotes. *Proteins*, **60**(4):670–8, Sep 2005.

[143] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, **26**(16):1781–802, Dec 2005.

[144] P. F. Predki, V. Agrawal, A. T. Brünger, and L. Regan. Amino-acid substitutions in a surface turn modulate protein stability. *Nat Struct Biol*, **3**(1):54–8, Jan 1996.

[145] E. Querido, P. Blanchette, Q. Yan, T. Kamura, M. Morrison, D. Boivin, W. G. Kaelin, R. C. Conaway, J. W. Conaway, and P. E. Branton. Degradation of p53 by adenovirus E4orf6 and E1B55K proteins occurs via a novel mechanism involving a Cullin-containing complex. *Genes Dev*, **15**(23):3104–17, Dec 2001.

[146] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[147] C. R. Robinson and R. T. Sauer. Striking stabilization of Arc repressor by an engineered disulfide bond. *Biochemistry*, **39**(40):12494–502, Oct 2000.

[148] P. M. Rudd, T. Elliott, P. Cresswell, I. A. Wilson, and R. A. Dwek. Glycosylation and the immune system. *Science*, **291**(5512):2370–6, Mar 2001.

[149] A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, **234**(3):779–815, Dec 1993.

[150] A. B. Sandler and G. Ketner. Adenovirus early region 4 is essential for normal stability of late nuclear RNAs. *J Virol*, **63**(2):624–30, Feb 1989.

[151] C. N. Scanlan, J. Offer, N. Zitzmann, and R. A. Dwek. Exploiting the defensive sugars of HIV-1 for drug and vaccine design. *Nature*, **446**(7139):1038–45, Apr 2007.

[152] M. Schaefer, C. Bartels, and M. Karplus. Solution conformations and thermodynamics of structured peptides: molecular dynamics simulation with an implicit solvation model. *J Mol Biol*, **284**(3):835–48, Dec 1998.

[153] M. Schaefer and M. Karplus. A comprehensive analytical treatment of continuum electrostatics. *The Journal of Physical Chemistry*, **100**(5):1578–99, 1996.

[154] G. Schiedner, S. Hertel, and S. Kochanek. Efficient transformation of primary human amniocytes by E1 functions of Ad5: generation of new cell lines for adenoviral vector production. *Hum Gene Ther*, **11**(15):2105–16, Oct 2000.

[155] T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, **18**(20):6097–100, Oct 1990.

[156] C. D. Schwieters, J. J. Kuszewski, N. Tjandra, and G. M. Clore. The Xplor-NIH NMR molecular structure determination package. *J Magn Reson*, **160**(1):65–73, Jan 2003.

[157] J.-S. Seeler and A. Dejean. Nuclear and unclear functions of SUMO. *Nat Rev Mol Cell Biol*, **4**(9):690–9, Sep 2003.

[158] D. J. Segel, A. L. Fink, K. O. Hodgson, and S. Doniach. Protein denaturation: a small-angle X-ray scattering study of the ensemble of unfolded states of cytochrome *c*. *Biochemistry*, **37**(36):12443–51, Sep 1998.

[159] A. Sexton, P. M. Drake, N. Mahmood, S. J. Harman, R. J. Shattock, and J. K.-C. Ma. Transgenic plant production of cyanovirin-N, an HIV microbicide. *FASEB J*, **20**(2):356–88, Feb 2006.

[160] S. Shahzad-ul-Hussan, M. Cai, and C. A. Bewley. Unprecedented glycosidase activity at a lectin carbohydrate-binding site exemplified by the cyanobacterial lectin MVL. *J Am Chem Soc*, **131**(45):16500–8, Nov 2009.

[161] S. Shahzad-ul-Hussan, E. Gustchina, R. Ghirlando, G. M. Clore, and C. A. Bewley. Solution structure of the monovalent lectin microvirin in complex with Manα(1-2)Man provides a basis for anti-HIV activity with low toxicity. *J Biol Chem*, **286**(23):20788–796, Jun 2011.

[162] T. H. Shen, H.-K. Lin, P. P. Scaglioni, T. M. Yung, and P. P. Pandolfi. The mechanisms of PML-nuclear body formation. *Mol Cell*, **24**(3):331–9, Nov 2006.

[163] S. R. Shenoy, L. G. Barrientos, D. M. Ratner, B. R. O'Keefe, P. H. Seeberger, A. M. Gronenborn, and M. R. Boyd. Multisite and multivalent binding between cyanovirin-N and branched oligomannosides: calorimetric and NMR characterization. *Chem Biol*, **9**(10):1109–18, Oct 2002.

[164] R. N. Shepard and D. A. Ornelles. E4orf3 is necessary for enhanced S-phase replication of cell cycle-restricted subgroup C adenoviruses. *J Virol*, **77**(15):8593–5, Aug 2003.

[165] R. N. Shepard and D. A. Ornelles. Diverse roles for E4orf3 at late times of infection revealed in an E1B 55-kilodalton protein mutant background. *J Virol*, **78**(18):9924–35, Sep 2004.

[166] S. B. Shuker, P. J. Hajduk, R. P. Meadows, and S. W. Fesik. Discovering high-affinity ligands for proteins: SAR by NMR. *Science*, **274**(5292):1531–4, Nov 1996.

[167] C. K. Smith, J. M. Withka, and L. Regan. A thermodynamic scale for the beta-sheet forming tendencies of the amino acids. *Biochemistry*, **33**(18):5510–17, May 1994.

[168] S.-Y. Sohn and P. Hearing. Adenovirus regulates SUMOylation of Mre11-Rad50-Nbs1 components through a paralog-specific mechanism. *Journal of Virology*, **86**(18):9656–65, Sep 2012.

[169] J. Song, Z. Zhang, W. Hu, and Y. Chen. Small ubiquitin-like modifier (SUMO) recognition of a SUMO binding motif: a reversal of the bound orientation. *J Biol Chem*, **280**(48):40122–9, Dec 2005.

[170] C. Soria, F. E. Estermann, K. C. Espantman, and C. C. O'Shea. Heterochromatin silencing of p53 target genes by a small viral protein. *Nature*, **466**(7310):1076–81, Aug 2010.

[171] S. Spector, M. Wang, S. A. Carp, J. Robblee, Z. S. Hendsch, R. Fairman, B. Tidor, and D. P. Raleigh. Rational modification of protein stability by the mutation of charged surface residues. *Biochemistry*, **39**(5):872–9, Feb 2000.

[172] J. Srinivasan, T. E. Cheatham, P. Cieplak, P. A. Kollman, and D. A. Case. Continuum solvent studies of the stability of DNA, RNA, and Phosphoramidate–DNA helices. *Journal of the American Chemical Society*, **120**(37):9401–9, 1998.

[173] T. H. Stracker, C. T. Carson, and M. D. Weitzman. Adenovirus oncoproteins inactivate the Mre11-Rad50-NBS1 DNA repair complex. *Nature*, **418**(6895):348–52, Jul 2002.

[174] T. H. Stracker, D. V. Lee, C. T. Carson, F. D. Araujo, D. A. Ornelles, and M. D. Weitzman. Serotype-specific reorganization of the Mre11 complex by adenoviral E4orf3 proteins. *J Virol*, **79**(11):6664–73, Jun 2005.

[175] V. Tiwari, S. Y. Shukla, and D. Shukla. A sugar binding protein cyanovirin-N blocks herpes simplex virus type-1 entry and cell fusion. *Antiviral Res*, **84**(1):67–75, Oct 2009.

[176] A. J. Ullman and P. Hearing. Cellular proteins PML and Daxx mediate an innate antiviral defense antagonized by the adenovirus E4 ORF3 protein. *J Virol*, **82**(15):7325–35, Aug 2008.

[177] A. J. Ullman, N. C. Reich, and P. Hearing. Adenovirus E4 ORF3 protein inhibits the interferon-mediated antiviral response. *J Virol*, **81**(9):4744–52, May 2007.

[178] A. G. van der Veen and H. L. Ploegh. Ubiquitin-like proteins. *Annu Rev Biochem*, **81**:323–57, Jan 2012.

[179] M. Vijayan and N. Chandra. Lectins. *Curr Opin Struc Biol*, **9**(6):707–14, Dec 1999.

[180] E. I. Vink, M. A. Yondola, K. Wu, and P. Hearing. Adenovirus E4-ORF3-dependent relocalization of TIF1α and TIF1γ relies on access to the coiled-coil motif. *Virology*, **422**(2):317–25, Jan 2012.

[181] C. D. Waldburger, J. F. Schildbach, and R. T. Sauer. Are buried salt bridges important for protein stability and conformational specificity? *Nat Struct Biol*, **2**(2):122–8, Feb 1995.

[182] L.-X. Wang, J. Ni, S. Singh, and H. Li. Binding of high-mannose-type oligosaccharides and synthetic oligomannose clusters to human antibody 2G12: implications for HIV-1 vaccine design. *Chem Biol*, **11**(1):127–34, Jan 2004.

[183] X. Wei, J. M. Decker, S. Wang, H. Hui, J. C. Kappes, X. Wu, J. F. Salazar-Gonzalez, M. G. Salazar, J. M. Kilby, M. S. Saag, N. L. Komarova, M. A. Nowak, B. H. Hahn, P. D. Kwong, and G. M. Shaw. Antibody neutralization and escape by HIV-1. *Nature*, **422**(6929):307–12, Mar 2003.

[184] M. D. Weiden and H. S. Ginsberg. Deletion of the E4 region of the genome produces adenovirus DNA concatemers. *Proc Natl Acad Sci USA*, **91**(1):153–7, Jan 1994.

[185] M. D. Weitzman, C. E. Lilley, and M. S. Chaurushiya. Genomes in conflict: maintaining genome integrity during virus infection. *Annu Rev Microbiol*, **64**:61–81, Jan 2010.

[186] M. D. Weitzman and D. A. Ornelles. Inactivating intracellular antiviral responses during adenovirus infection. *Oncogene*, **24**(52):7686–96, Nov 2005.

[187] A. Whitty. Cooperativity and biological complexity. *Nat Chem Biol*, **4**(8):435–9, Aug 2008.

[188] D. C. Williams, J. Y. Lee, M. Cai, C. A. Bewley, and G. M. Clore. Crystal structures of the HIV-1 inhibitory cyanobacterial protein MVL free and bound to $Man_3GlcNAc_2$: structural basis for specificity and high-affinity binding to the core pentasaccharide from N-linked oligomannoside. *J Biol Chem*, **280**(32):29269–76, Aug 2005.

[189] J. R. Williamson. Cooperativity in macromolecular assembly. *Nat Chem Biol*, **4**(8):458–65, Aug 2008.

[190] P. Wimmer, P. Blanchette, S. Schreiner, W. Ching, P. Groitl, J. Berscheminski, P. E. Branton, H. Will, and T. Dobner. Cross-talk between phosphorylation and SUMOylation regulates transforming activities of an adenoviral oncoprotein. *Oncogene*, May 2012.

[191] P. Wimmer, S. Schreiner, R. D. Everett, H. Sirma, P. Groitl, and T. Dobner. SUMO modification of E1B-55K oncoprotein regulates isoform-specific binding to the tumour suppressor protein PML. *Oncogene*, **29**(40):5511–22, Oct 2010.

[192] P. Wimmer, S. Schreiner, and T. Dobner. Human pathogens and the host cell SUMOylation system. *J Virol*, **86**(2):642–54, Jan 2012.

[193] J. L. Woo and A. J. Berk. Adenovirus ubiquitin-protein ligase stimulates viral late mRNA nuclear export. *J Virol*, **81**(2):575–87, Jan 2007.

[194] J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J Mol Biol*, **285**(4):1735–47, Jan 1999.

[195] M. R. Wormald, A. J. Petrescu, Y.-L. Pao, A. Glithero, T. Elliott, and R. A. Dwek. Conformational studies of oligosaccharides and glycopeptides: complementarity of NMR, X-ray crystallography, and molecular modelling. *Chem Rev*, **102**(2):371–86, Feb 2002.

[196] M. Yamaguchi, T. Ogawa, K. Muramoto, Y. Kamio, M. Jimbo, and H. Kamiya. Isolation and characterization of a mannan-binding lectin from the freshwater cyanobacterium (blue-green algae) *Microcystis viridis*. *Biochem Biophys Res Commun*, **265**(3):703–8, Nov 1999.

[197] F. Yang, C. A. Bewley, J. M. Louis, K. R. Gustafson, M. R. Boyd, A. M. Gronenborn, G. M. Clore, and A. Wlodawer. Crystal structure of cyanovirin-N, a potent HIV-inactivating protein, shows unexpected domain swapping. *J Mol Biol*, **288**(3):403–12, May 1999.

[198] M. A. Yondola and P. Hearing. The adenovirus E4 ORF3 protein binds and reorganizes the TRIM family member transcriptional intermediary factor 1 alpha. *J Virol*, **81**(8):4264–71, Apr 2007.

[199] H. X. Zhou, R. H. Hoess, and W. F. DeGrado. In vitro evolution of thermodynamically stable turns. *Nat Struct Biol*, **3**(5):446–51, May 1996.

[200] N. E. Ziółkowska, B. R. O'Keefe, T. Mori, C. Zhu, B. Giomarelli, F. Vojdani, K. E. Palmer, J. B. McMahon, and A. Wlodawer. Domain-swapped structure of the potent antiviral protein griffithsin and its mode of carbohydrate binding. *Structure*, **14**(7):1127–35, Jul 2006.