Inclusion Body Formation by Mutants of the Tenth Human Fibronectin Type III Domain

by

Kyle Trainor

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Chemistry

Waterloo, Ontario, Canada, 2015

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Inclusion bodies (IBs) are intracellular, insoluble protein aggregates, commonly observed when a protein of interest is expressed at high concentrations in a bacterial cell-based expression system. The molecular determinants of IB formation are poorly understood, and are of both fundamental and biotechnological significance.

The stability, folding, and structure of the tenth human fibronectin type III domain (^{10}Fn3) have been studied previously, making it an attractive model system to investigate IB formation. A library of ^{10}Fn3 mutants was provided by Bristol-Myers Squibb; 31 of these mutants were expressed in *Escherichia coli* and analyzed. The percentage of the expressed protein found within IBs was quantified at different expression time points using densitometric analysis of soluble and inclusion body (insoluble) cell lysate fractions separated by centrifugation and subjected to polyacrylamide gel electrophoresis. Although most of these mutants differ from each other in only 3 amino acid positions, all found within a single flexible loop of the protein, the extent of IB formation varies greatly.

This data set was used to test the performance of a variety of amino acid sequence-based protein aggregation prediction methods. Several of these methods produced predictions that correlate moderately well with the IB formation data ($R^2 > 0.6$), suggesting that while the intrinsic aggregation propensity of sequence segments strongly influences IB formation, other factors are also relevant. We hypothesized that improved predictions might be made possible by the consideration of additional structural context, i.e. aggregation-prone sequence segment exposure.

Thermodynamic stabilities determined using differential scanning calorimetry correlate poorly with IB formation; all of the mutants are sufficiently stable that no significant fraction of protein is likely to be denatured at equilibrium. To describe the variable structure of the flexible loop in which the mutant sequences differ, ensembles of homology models were constructed. IB formation was found to correlate with the ensemble average energy scores of the homology models. The ensemble average scores may capture subtle shifts in the energetic bias toward native structure that restricts the exposure of aggregation-prone sequence segments. A linear combination of sequence-based aggregation predictions and ensemble average homology model scores correlates much better with IB formation ($R^2 > 0.8$) than either parameter does individually.

Acknowledgements

I would like to thank my supervisor, Dr. Elizabeth Meiering, and all members of the Meiering lab, past and present; a supportive and knowledgeable group. Special thanks to Zachary Gingras, Allen Chiu, and Cicely Shillingford for their help with this project.

This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca) and Compute/Calcul Canada.

Dedication

To my parents, for encouraging independent thought and higher education, and to Melissa, for supporting my mad science dreams.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Fundamentally, proteins are linear chains of amino acids joined by covalent (peptide) bonds. The physicochemical properties of amino acid side chains vary widely, endowing each different sequence of amino acids with a unique character. For all but the shortest polypeptides, the number of possible conformations is vast [1], yet in most cases, evolved proteins are observed to fold into a specific, low-energy three-dimensional structure (the native state) on a biologically accessible timescale. The sequences of evolved proteins are therefore thought to be under selective pressure that favours those capable of an expeditious transition from an unfolded state to a stable (low energy) native fold. Within the context of the energy landscape theory of protein folding [2], this concept is described by the "folding funnel" hypothesis (Figure 1.1) [3].

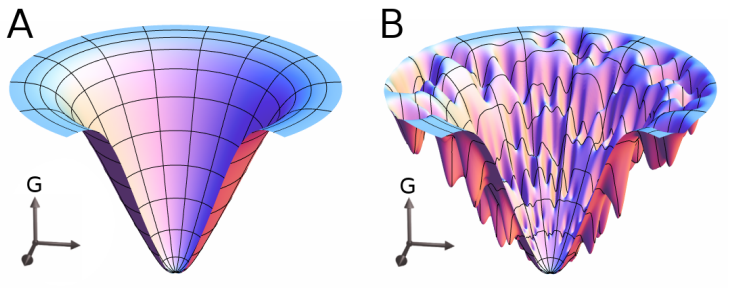

Figure 1.1: Free energy landscapes with Gibbs energy on the vertical axis, and a two-dimensional projection of conformational space on the horizontal axes; A: idealized folding funnel, B: rugged funnel with local minima. Figure produced using a Mathematica™ notebook provided by the Oas lab (http://www.oaslab.com).

Though the "unfolded state" is often referred to in the singular, the term is understood to imply an ensemble of diverse conformations having in common a lack of well-defined structure. In the absence of chemical denaturants, completely unfolded polypeptides are unlikely to persist at physiologically relevant temperatures; partially synthesized polypeptides can begin to fold while still attached to the ribosome [4]. The rate of folding is determined by the height of one or more energy barriers, each corresponding to a high-energy transition state (TS) structure that must be adopted before a conformation in the low-energy native ensemble (NE) can be attained (Figure 1.2). Before the rate-limiting energy barriers are surmounted, the hydrophobic effect [5] may drive the collapse of unfolded polypeptides into an ensemble of compact states [6, 7] in which stabilizing secondary structure can be found [8]. We will adopt the convention of referring to these compact, non-native conformations as the denatured state ensemble (DSE), in order to distinguish them from completely unfolded polypeptides.

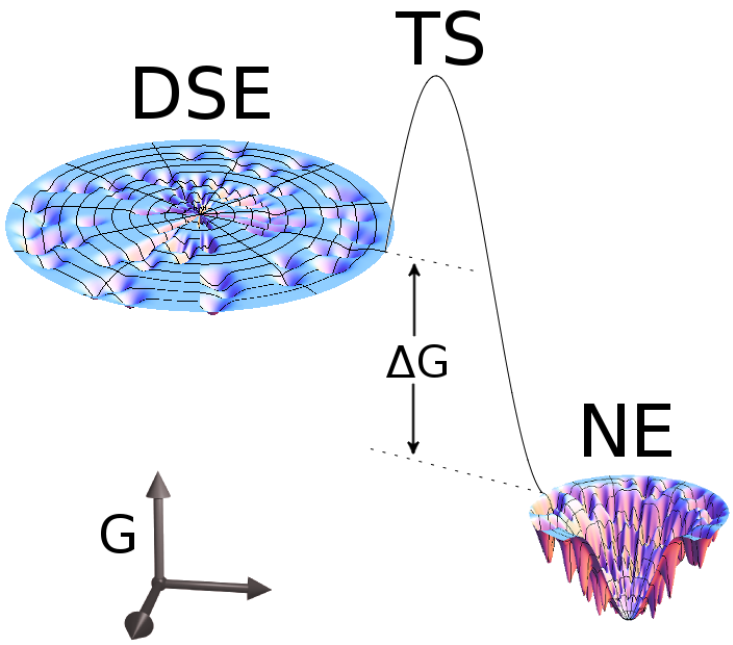

Figure 1.2: A free-energy diagram showing a path from the denatured state ensemble (DSE) to the native ensemble (NE) through a rate-limiting transition state (TS). There may also be an ensemble of transition state structures (omitted from this illustration for clarity). Figure produced using a Mathematica™ notebook provided by the Oas lab (http://www.oaslab.com).

Similarly, the native "state" is best envisioned as an ensemble of related conforma-

tions that define a low-energy basin in the energy landscape. The dynamic nature of native protein structure has been explored extensively using experimental methods such as hydrogen-deuterium (H-D) exchange [9, 10], and relaxation dispersion NMR spectroscopy [11]. High-energy "excited" states (sometimes relevant to protein function) have been detected and characterized [12, 13].

The selective pressure for thermodynamically stable proteins that fold quickly is tempered by competing (e.g. functional) constraints on protein sequence and structure, and deviations from ideal folding behaviour and stability may also result from random mutations. This may manifest as a rugged folding funnel with many local minima (Figure 1.1B), higher rate-limiting energy barriers (slower folding), a smaller energy difference between the NE and the DSE (lower global stability), or a greater diversity of structures in the NE (local openings).

The exposure of segments that would ideally be sheltered from intermolecular association (in the native structure) can lead to protein aggregation [14]. Unlike monomer folding (which is unimolecular), aggregation is a minimally bimolecular, concentration-dependent process. Protein aggregation is commonly encountered in the field of biotechnology when high concentrations of proteins are produced in an expression system. In a cell-based expression system such as *Escherichia coli* (*E. coli*), these aggregates are referred to as inclusion bodies.

1.1 Inclusion Body Formation and Structure

The term "inclusion bodies" broadly encompasses all intracellular aggregates. Among the best studied inclusion bodies (IBs) are those formed upon overexpression of a heterologous gene in *E. coli*. An indispensable workhorse of biotechnology since the birth of the field [15], *E. coli* remains one of the most widely used cell-based protein expression systems due to its status as a well-characterized model organism [16], and because it is relatively easy to grow cultures at high cell density [17]. A better understanding of the determinants of IB formation would be of great practical value; though their homogeneous composition simplifies the purification process (80-95% of a typical IB is made up of the overexpressed protein [18]), the constituent proteins must be refolded. The refolding process varies greatly in difficulty and efficiency from protein to protein [19]. The desirability of soluble vs. insoluble (inclusion body) expression thus depends on the context.

Even considering only IBs formed during heterologous protein overexpression in *E. coli*, many types of IBs can be differentiated on the basis of location (cytoplasm or periplasmic

space), morphology, and ease of solubilization [19, 20]. The secondary structure inside IBs may vary from near-native to distinctly non-native [21, 22, 23, 24, 25]. Further sub-categorization of bacterial IBs is possible if the overexpression of multi-domain proteins, oligomeric proteins, and intrinsically disordered polypeptides are considered separately from that of single-domain proteins that fold to a stable, monomeric state under normal physiological conditions.

To date, attempts to find correlations with predictive value between IB formation and protein sequence, stability, or simple physicochemical properties have yielded almost as many exceptions as rules. Software designed to predict soluble expression has been developed by training machine learning algorithms on data from tens of thousands of proteins; the highest accuracy reported from amongst these methods is approximately 75% [26]. This is an impressive result, given the large number of confounding variables present in an environment as complex and dynamic as the interior of a living cell. Proteins and other macromolecules are present at concentrations of up to 300-400 g/L [27], which endows the cytoplasm with a gel-like consistency, and gives rise to macromolecular crowding effects that may promote protein-protein association [28]. Also, many of these proteins are far from passive bystanders: proteases [29, 30], chaperones [31, 32] and molecular machines responsible for the transport of small aggregates to the poles of *E. coli* cells [33] are but a few examples of active intracellular macromolecules that may impact the expressed protein. Growth conditions such as temperature [24] and induction level [23] have also been shown to affect IB structure, and the partitioning of the expressed protein between the soluble and insoluble cellular fractions. Many aspects of intracellular complexity have been reviewed elsewhere [29, 30, 31, 32].

Relative to unfolded or denatured states, both native folding and aggregate formation are likely to be energetically favoured under physiological conditions [34]; if the barriers to denaturation and disaggregation are sufficiently high, the fate of denatured proteins under folding conditions may be determined by a kinetic competition [35, 36]. Historically, it was widely believed that interactions between the exposed hydrophobic side chains of unstructured proteins were primarily responsible for the formation of bacterial inclusion bodies [37]. The burial of hydrophobic residues remains a major factor in the stability of any water-solvated protein conformation, but more recent studies have emphasized intermolecular association mediated by aggregation-prone stretches of self-complementary sequence [38, 39, 40]. The specificity of the interactions between such sequence segments appears to be high; even when two IB-forming proteins are expressed simultaneously, they may not appreciably co-aggregate (true co-aggregation can be distinguished from co-localization by Förster resonance energy transfer) [38, 41]. In theory, this self-complementarity could be attributed solely to hydrophobic side chain burial, but the enrichment of β-sheet structure

frequently observed within IBs [23, 24, 37] suggests that hydrogen bonding and other polar interactions also contribute to inclusion body formation and stability.

We will first review mechanisms of intermolecular association that may be at work within IBs, and then consider aggregation-prone sequence segment exposure during the competition between aggregation and native protein folding *in vivo*.

1.1.1 Mechanisms of Intermolecular Association

There are three common types of non-covalent intermolecular interactions that may be particularly pertinent to the stabilization of protein-protein associations: hydrophobic side chain burial, predominantly polar contacts, and intermolecular β-sheet (Table 1.1).

Table 1.1: **Comparison of different types of intermolecular contacts.**

Intermolecular Interaction	Characteristics
Hydrophobic Side Chain Burial	Burial of hydrophobic residues in an intermolecular interface.
Polar Contacts	Numerous interactions between polar residues, possibly distributed over multiple contact patches.
Intermolecular β-sheet	Secondary structure-promoting hydrogen bonds, often combined with burial of hydrophobic side chains.

Intermolecular association driven by the burial of complementary hydrophobic surfaces has long been known to be a mechanism of protein complex formation [42]. The surfaces in question need not be exclusively hydrophobic; though the hydrophobic effect may be dominant, polar interactions at the interface can also play a stabilizing role [43], and either a single, contiguous hydrophobic patch or a number of smaller patches are possible [44]. The feature that differentiates this type of intermolecular association is the mean hydrophobicity of residues in the interface, which is greater than that of residues on the solvent-exposed exterior of the complex [45]. Hydrophobic surfaces that are safely buried in the monomeric native state may be exposed in fully or partially unstructured conformations. Aggregates stabilized primarily by non-specific hydrophobic interactions (i.e. between surfaces or segments that are not complementary) are also possible. However, this type of association does not provide a satisfactory explanation for the homogeneous composition of IBs [38, 41].

At the opposite end of the hydrophobicity spectrum, polar contacts, such as those commonly found in protein crystals, are formed by the burial of predominantly polar surfaces in an intermolecular interface. Crystals upon which X-ray crystallography is performed are typically composed of natively folded proteins. Under normal physiological conditions, these proteins are soluble, and the amino acid composition of the crystal-packing interface is often virtually indistinguishable from that of the solvent-exposed surface [46]. However, just as polar residues may be found in predominantly hydrophobic surfaces, non-polar ones can be found in crystal-like contacts; the task of distinguishing "biological" from "crystal-packing" interfaces can be non-trivial [47]. We will define polar contacts as those in which the intermolecular interaction is predominantly mediated by favourable interactions between polar and charged amino acid side chains, rather than the hydrophobic effect [48].

The intermolecular β-sheet type of protein-protein association frequently involves not only polar intermolecular contacts (hydrogen bonds), but also sequestration of hydrophobic amino acid side chains. It is unique in that segments from multiple polypeptides collectively form highly stabilizing β-structure. This type of intermolecular association is commonly observed in domain-swapping oligomerization [49], and forms the basis of the amyloid fibril spine [50]. Evidence has been found of non-native β-structure in inclusion bodies [21, 23, 37].

All three types of intermolecular contacts listed in Table 1.1 are possible, even likely, to be found within inclusion bodies, given the very high local concentration of both folded and fully or partially unstructured polypeptides that may exist during overexpression. The specificity of any of the three types of intermolecular association mechanisms may be sufficient to account for the homogeneous composition of IBs [18, 38, 41], but only intermolecular β-sheet formation explains the frequently observed increase in β-structure [23, 24, 37].

1.1.2 Amyloid-Like Stabilization Mechanisms

The difficulty of solubilizing proteins from IBs varies greatly; however, with rare exceptions [51], simple resuspension in an appropriate buffer is not sufficient to solubilize an appreciable fraction of IB protein content [19]. Crystal-like polar contacts may be established when the concentration of an overexpressed protein exceeds its solubility limit, but aggregates formed in this fashion are easily redissolved [52].

In some cases, low concentrations of detergents [53, 54] or other disaggregating agents such as L-arginine [55, 56] can be sufficient to solubilize IB proteins. More commonly, very

high concentrations of a denaturant or strong detergent are required [19]. The non-native β-structure content and high stability typical of IBs in these latter cases are characteristics shared with fibrillar amyloid aggregates [39, 40, 57].

Amyloid aggregates feature intermolecular interactions between β-strands in a highly extended, tightly packed structure [50, 58, 59, 60]. The amorphous morphology and imperfect intermolecular β-sheet alignment of IBs [39] contrast with the properties of amyloid fibrils, but the similarities are otherwise compelling (Table 1.2). It is also worth noting that the digestion of apparently amorphous bacterial inclusion bodies by Proteinase K has been observed to leave behind a tangle of resistant fibril-like structures [38], and solid-state nuclear magnetic resonance (NMR) spectra show that HET-s(218-289) adopts the same structure in IBs (produced in *E. coli*) as it does in fibrils (produced *in vitro*) [61].

Table 1.2: **Comparison of amyloid fibrils and inclusion bodies.**

	Amyloid Fibrils	**Inclusion Bodies**
Proposed Stabilizing Intermolecular Interaction	Cross-β spine, with β-strands oriented perpendicular to the fibre axis [58, 59, 60].	Intermolecular β-sheet structure formed by amyloid-like, but imperfectly aligned, extended β-strands [39, 57].
Dye Binding	Amyloid-specific dyes are bound. CR birefringence is observed due to the presence of aligned, repetitive structure [62, 63].	Amyloid-specific dyes are bound. CR birefringence may be observed [39, 40, 57].
X-ray Diffraction Pattern	A sharp reflection at approximately 4.7Å and a diffuse reflection at approximately 10Å [64].	A sharp reflection at approximately 4.7Å and a diffuse reflection at approximately 10Å [39, 65].
Morphology	Highly ordered [58, 59, 60].	Apparently amorphous [38] .
Solubility	Insoluble [64].	Insoluble [37].
Homogeneity	High [58, 59, 60].	High [18].

As a consequence of these striking similarities, it has been proposed that the formation of amyloid-like intermolecular β-sheet structure is the principal aggregate-stabilizing mechanism at work within IBs [57]. According to one recent analysis, the fraction of open reading frames (ORFs) in the *E. coli*, *S. cerevisiae*, and *H. sapiens* genomes containing at least one amyloidogenic sequence segment is over 99% [66]. Subsequent examination

of structures from the Protein Data Bank (PDB) showed that in natively folded proteins such amyloidogenic segments are normally inaccessible (buried within the hydrophobic core) or adopt conformations incompatible with intermolecular β-sheet formation; however, these normally inaccessible segments may be transiently exposed in fully or partially unstructured conformations populated during recombinant overexpression.

The proposal that IBs are stabilized by amyloid-like structure has not met with universal agreement. It has been argued that there are important distinctions between amyloid and the intermolecular β-structure in which amorphous aggregates are enriched [67]. We have adopted the convention of referring to intermolecular β-sheet structure formed by extended (relative to typical native structure) β-strands as "amyloid-like", despite these possible distinctions.

1.1.3 Inclusion Body Structure

Spectroscopic evidence of native-like secondary structure within IBs [21, 68], the existence of IBs containing functional polypeptides [25, 69], and the strong correlation observed between the predicted rate of aggregation and the degree of native-like structure in GFP-Aβ42 fusion protein variants overexpressed in *E. coli* [70] all support the hypothesis that the competition between aggregation and native folding is not a simple winner-takes-all mechanism. In this section, we will consider aggregation-prone sequence segment exposure in various ensembles of conformations, and attempt to relate it to the varying degrees of native-like structure possible within IBs (Table 1.3).

Aggregation-prone segments may be exposed in the ensembles of fully unfolded, denatured state, intermediate, and native conformations. The ensemble that the constituent proteins of an IB are drawn from is a major factor determining the secondary and tertiary structure of those constituents, though post-aggregation folding/unfolding or structural changes cannot be ruled out. Because fully unfolded conformations are unlikely to be populated under physiologically relevant conditions (see below), and many proteins that have no detectable intermediates still form IBs [40], we will begin with a discussion of aggregation from the denatured state ensemble (DSE) and the native ensemble (NE). Aggregation from intermediate ensembles will then be considered.

Aggregation from the Denatured State Ensemble

It is fundamental to energy landscape theory that denatured proteins populate an ensemble of conformations [2]. Though ambiguous, there is evidence for some polypeptides of residual

Table 1.3: **Structure in inclusion bodies.**

Secondary and Tertiary Structure	Source of Constituent Proteins	Likely IB Characteristics
Non-native	Unfolded ensemble	Minimal secondary/tertiary structure except that which may be formed intermolecular contacts.
Variable	Intermediate or denatured state ensemble	Degree of native-like structure may vary widely.
Native-like	Native ensemble	Associations between locally unfolded aggregation prone sequence segments; runaway domain swapping possible.

secondary structure even in high levels of chemical denaturants [71]. More relevant to IB formation, in the absence of denaturing influences, i.e. under conditions that promote folding and/or aggregation, it is clear from the basic principles of statistical mechanics that occupancy of a truly unfolded state is exceedingly unlikely; proteins will tend to populate a DSE featuring relatively compact, low-energy conformations.

By definition, the rate-limiting transition energy barrier is much larger than the energy barriers separating the conformations in the fully unfolded and denatured state ensembles (Figure 1.2). Because different DSE conformations are separated by small energy barriers (relative to the rate-limiting barrier), the DSE will achieve a state of pseudo-equilibrium. The probability P of a given conformation i being populated in the DSE at temperature T is therefore determined by its energy E_i [72], as in equations 1.1 and 1.2.

$$DSE \xrightarrow{\text{Rate-Limiting Energy Barrier}} NE \tag{1.1}$$

$$P(E_i) \propto e^{-E_i/kT} \tag{1.2}$$

For the majority of proteins, the hydrophobic effect will favour compact conformations in the DSE [73, 74, 75], and experiments have confirmed that these conformations can include native-like structure [76, 77, 78]. Where this native-like structure exists, it may restrict aggregation; residues that are structured in the transition state of human muscle

acylphosphatase (AcP) appear to correspond with those that are protected (i.e. unavailable to participate in aggregation) in the DSE [36].

The Transition State and Structure in the DSE

The results of folding and aggregation experiments at a fixed protein concentration have been combined with a ϕ-value analysis to demonstrate that mutations in regions that are highly structured in the transition state of AcP can affect the folding rate independently of the aggregation propensity. Conversely, mutations in regions that are less structured (more solvent-exposed) in the transition state can affect the aggregation propensity independently of the folding rate [36]. In this case, and perhaps others like it, the degree of native-like structure in the DSE appears to be closely related to the transition state structure; regions that are unstructured in the transition state are also more likely to be unstructured (and available for aggregation) in the DSE. This has implications not only for aggregation prediction (Chapter 4), but also for IB structure (Table 1.3).

Aggregation from the Native-Like Ensemble

Except where limited solubility of the native state is an issue, polypeptides are generally thought to aggregate from a fully or partially unstructured state [37], yet many proteins that fold on a millisecond timescale nevertheless form inclusion bodies [40], and pulse-chase radiolabelling experiments have shown that it is possible for folded, soluble proteins to migrate to the insoluble fraction, even hours after translation [79]. How can these observations be reconciled? Natively folded proteins populate an ensemble of conformations relatively close to a well-defined average structure, rather than a single, rigid native state [80]. Flexible proteins have a particularly wide range of conformations (separated by low energy barriers) available to them [81], but evidence suggests that even inflexible proteins that fold in an apparently two-state manner may populate an ensemble of conformations in which locally, rather than globally, unfolded states are adopted (Figure 1.2) [82]. Apparent two-state folding mechanisms may sometimes be observed as a result of experimental limitations that prevent the detection of short-lived intermediates and native state fluctuations [83]. Though it is frequently both justifiable and useful to approximate folding as two-state, in the context of aggregation we must consider the possibility that aggregation-prone sequence segments may be exposed in post-transition state conformations, particularly under non-equilibrium (e.g. during folding from a denatured state) or nonstandard (e.g. elevated temperature) conditions [84, 85]. Indeed, there are well-documented cases of aggregation

(*in vitro*) from the native-like ensemble (NE) in which no transition across the cooperative unfolding energy barrier was required [86, 87, 88].

One relatively common form of aggregation from the NE is runaway domain swapping. In this case, locally unfolded states permit the formation of domain-swapped assemblies; if the domain swapping is open-ended, the result can be the formation of large aggregates (Figure 1.3). Domain swapping restricts the translational and rotational degrees of freedom of the protein, which is entropically unfavourable [49]; to compensate, the conformation adopted by proteins in the assembly must either relieve strain present in the monomer, or form stabilizing structure in the intermolecular interface. In order for domain swapping to occur, some segment must adopt the role of a "hinge loop" linking the swapped domain to the rest of the protein. If hinge loop strain in the closed monomeric form of the protein is relieved in a more extended form, the entropic penalty may be overcome. Strain induced by mutational hinge loop shortening has been observed to result in dimer formation [89]. More relevant to the study of IBs, the formation of domain-swapped aggregates is facilitated by longer, more flexible hinge loops capable of favourable self-association (β-sheet) in their extended form (Figure 1.3) [90].

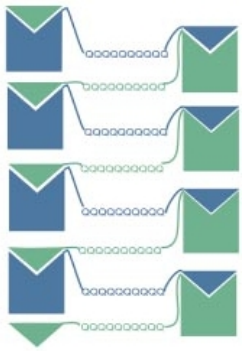

Figure 1.3: An illustration of open-ended, or "runaway" domain swapping. Blue and green subunits swap domains, and the hinge loops form stabilizing β-sheet structure. Reproduced from [90].

Aggregation from Intermediate Ensembles

An ensemble is defined by barriers in the energy landscape, but the landscape may contain many small barriers (Figure 1.1B). It is difficult to define exactly how large the energy barriers must be before the conformations that they isolate should be considered an ensemble with potential relevance to IB formation (i.e. sufficiently populated, either transiently or

at equilibrium). The DSE could be considered intermediate between the unfolded and native ensembles, but we concluded above that unfolded states are unlikely to be populated under physiological conditions. Our definition of the NE encompasses diverse states, some of which could be recast as intermediates. Intermediate ensembles not subsumed by our broadly defined DSE and NE are possible. The probability that partially folded structures from such ensembles contribute to IB formation is determined by the heights of the surrounding energy barriers.

1.2 Fibronectin Type III Domains

The model system we have used to study IB formation is based upon the tenth human fibronectin type III domain (^{10}Fn3) (Figure 1.4). Fibronectin is a multidomain glycoprotein, composed of repeats of three distinct domain types (I, II, and III), that is involved in cellular interactions with the extracellular matrix (ECM) [91]. The β-sandwich fold adopted by the fibronectin type III domains (Fn3), considered to be part of the immunoglobulin superfamily, is by no means limited to fibronectin; by one estimate, this fold can be found in approximately 2% of all animal proteins [92].

In its role as part of the ECM, fibronectin is assembled into elastic fibrils [93]. Studies of Chinese hamster ovary cells expressing green fluorescent protein-fibronectin chimeras have demonstrated that fibronectin fibrils can be extended up to four times their equilibrium length *in vivo* [94]. The extensibility of structural proteins such as fibronectin is a consequence of reversible unfolding of their constituent domains [95, 96]. This unfolding may serve to expose "cryptic" binding sites that promote the association of individual fibronectin molecules into fibrils [97, 98, 99]. The ^{10}Fn3 domain is thought to be among the first to unfold due to its low mechanical stability [100], and there is evidence that one of the cryptic binding sites is located in strand B of ^{10}Fn3 [101].

1.2.1 Mechanical Unfolding

Because the wild-type ^{10}Fn3 domain is suspected to mediate fibronectin fibrillogenesis by unfolding under mechanical stress to reveal cryptic binding sites [101], ^{10}Fn3 unfolding has been studied extensively using both experimental and computational methods. Single-molecule force spectroscopy experiments [102] and steered molecular dynamics simulations [103, 104, 105] in which a tensile force was applied between the N and C-termini (similar to forces that might be applied to a ^{10}Fn3 domain through the ECM) independently concluded

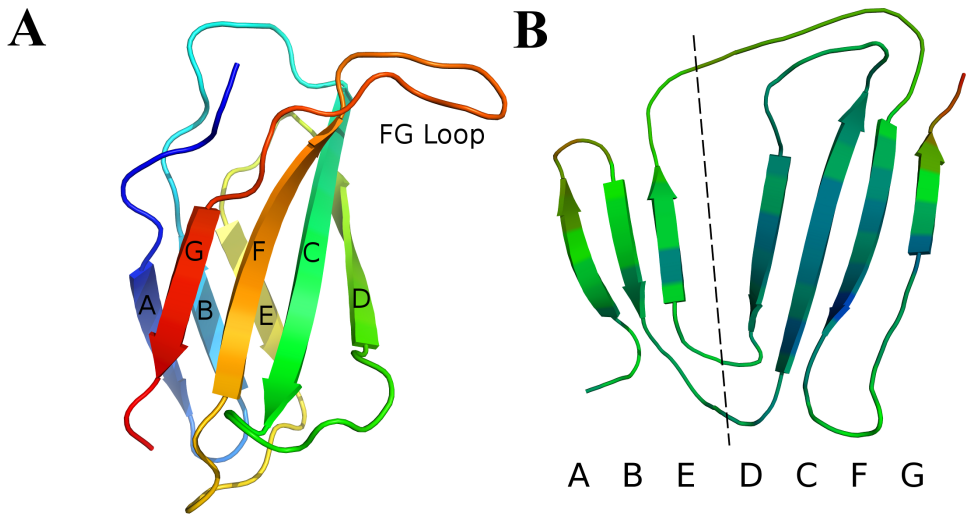

Figure 1.4: Cartoon representations of wild-type ^{10}Fn3 (derived from the crystal structure with PDB identifier 1FNF). A: coloured from N-terminus (blue) to C-terminus (red). B: flat representation coloured by B-factor, from low (blue) to high (red); loops not to scale. Dashed line divides the two halves of the β-sandwich. Figure produced using PyMOL and GIMP.

that at least two different unfolding intermediates are observed (as part of two distinct pathways): one in which β-strands A and B become detached and solvent-exposed, and another in which β-strand G becomes detached.

The application of a tensile force between the N-terminus and the integrin-binding FG loop (a loading pattern that is physiologically relevant if cell-traction forces are responsible for ^{10}Fn3 unfolding) has also been simulated. These simulations consistently feature an unfolding intermediate in which β-strand A is detached [105].

1.2.2 Structural Dynamics

Like fibronectin, tenascin is a large, multi-domain protein found in the ECM. Approximately half of the domains in each of these proteins are classified as Fn3 folds [106]. A domain homologous to ^{10}Fn3, the third fibronectin type III domain of human tenascin (^{3}TnFn3), has also been shown to unfold under mechanical stress through a force-stabilized intermediate in which the A and/or G β-strands may be unstructured [107].

Spin relaxation NMR experiments have revealed conformational dynamics of ^{3}TnFn3

that include the collective motion of β-sheets [108]. Residues with correlation times and dispersion amplitudes indicative of conformational exchange on a microsecond/millisecond timescale were also found to have unusually large B-factors in the crystal structure (PDB ID: 1TEN) [109]. Some of the corresponding β-strands of ^{10}Fn3 (crystal structure PDB ID: 1FNF) include atoms with similarly large B-factors (Figure 1.4B) [110].

H-D exchange experiments performed on ^{10}Fn3 demonstrated that a large fraction of residues in the A and G strands are not protected against exchange; the corresponding residues in ^{3}TnFn3 are measurably protected, despite its lower global stability, which suggests that these regions of ^{10}Fn3 are particularly dynamic [111].

1.2.3 Folding and Transition State Structure

The folding of ^{10}Fn3 has been modelled as a three-state transition, with a folding intermediate apparent at low concentrations of denaturant [112]. Detailed ϕ-value analysis of ^{10}Fn3 has shown that β-strands C, D, E, and F are significantly structured in the transition state (though all of the ϕ-values are fractional) [113] (Figure 1.5). The corresponding β-strands are also highly structured in the transition state of ^{3}TnFn3 [114], and it has been proposed that these two proteins share a common folding nucleus [113].

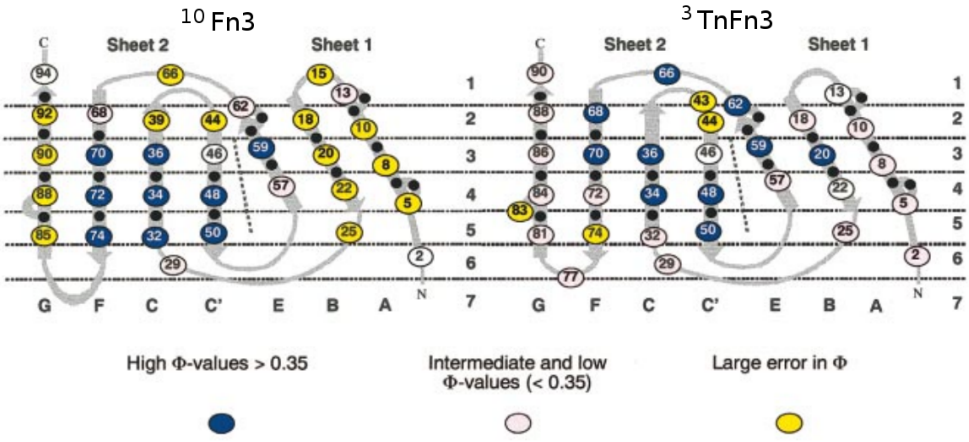

Figure 1.5: The results of ϕ-value analyses of wild-type ^{10}Fn3 (left) and ^{3}TnFn3 (right). Regions thought to be structured in the transition state are indicated by blue ovals. Reproduced from [113].

The A, B, and G strands are unambiguously less structured than the folding nucleus in the ^{3}TnFn3 transition state, but anomalously small changes in ^{10}Fn3 ΔG (Figure 1.2)

upon mutation of residues in these strands (even those that appear to be deeply buried in the hydrophobic core of the native structure) complicate interpretation of the associated ϕ-values. This unusual accommodation of mutations without loss of stability has been attributed to the dynamic character of the native structure (Section 1.2.2) [111].

1.2.4 Adnectins™

The structural similarity of Fn3 to the immunoglobulin fold (including solvent accessible loops resembling the V_H complementarity-determining regions H1, H2, and H3 of immunoglobulin), combined with favourable characteristics such as high thermostability, solubility, and expression level, as well as the absence of disulfide bonds or free cysteine residues [115], has lead to the development of antibody mimics for therapeutic applications based upon ^{10}Fn3 [116, 117]. The Adnectins™, a family of such proteins with mutated BC, DE, and FG loops (Figure 1.4) [118], are among the earliest developed and most advanced engineered target-binding proteins [115].

We have studied the IB formation, stability, and aggregation propensity of a group of Adnectins™ (Appendix A) that differ from each other almost exclusively in the FG loop (the two exceptions, 5898_C02 and 5898_F01, have additional mutations in the DE loop). The FG loops of these Adnectins™ are 6 residues in length (4 residues shorter than the corresponding wild-type loop), and the sequences vary only in the first, second, and sixth positions. These small variations result in dramatically different IB formation propensities when the Adnectins™ are expressed in *E. coli* (Chapter 2), despite the fact that the native folds of all mutants characterized to date appear to be both stable and soluble (Chapter 3).

Chapter 2

Quantification of Inclusion Body Formation

2.1 Introduction

Insoluble intracellular aggregates formed upon overexpression of a heterologous protein in a bacterial expression system are commonly referred to as inclusion bodies (Chapter 1). Bacterial cell lysate can contain both insoluble IBs and natively folded, soluble protein. These two forms are easily separated by centrifugation, and the percentage of expressed protein found in the insoluble fraction under a particular set of growth conditions can be quantified by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and densitometry, and taken as a measure of the IB formation propensity of that protein [119, 120].

The Adnectins™ included in this study (Appendix A) were chosen on the basis of sequence similarity. With the exceptions of Adnectins™ 5898_C01 and 5898_F01, all of the sequences are identical outside of the FG loop. Also, all of the FG loop sequences have in common arginine, aspartate, and tyrosine in the third, fourth, and fifth FG loop positions, leaving variation only in the first, second, and sixth positions. The Adnectins™ are known to display varying IB formation propensities; even a single mutation in a flexible loop can substantially shift the soluble vs. IB balance [121]. By selecting a group of very closely related sequences, we hoped to reduce the number of uncontrolled variables, and simplify the task of identifying some of the determinants of IB formation.

2.2 Methods

2.2.1 Protein Expression

The mutants were expressed in BL21 (DE3) cells (purchased from Edge Biosystems) containing the pLysS plasmid, which reduces basal expression of recombinant genes under the control of the T7 promoter by producing low levels of T7 lysozyme (a natural inhibitor of T7 RNA polymerase). Glycerol stocks of such cells, transformed with pET-9d vectors into which the Adnectins had been cloned at the NcoI and BamHI sites, were provided by D. Lipovšek (Adnexus, a division of Bristol-Myers Squibb); these stocks were maintained at -80°C.

Cells from the glycerol stocks were streaked onto agar plates containing 50 µg/mL kanamycin and 34 µg/mL chloramphenicol, and grown overnight at 37°C. A single colony was selected and transferred to a 50 mL conical tube containing 15 mL of sterile LB broth [122] with 50 µg/mL kanamycin and 34 µg/mL chloramphenicol, and grown overnight at 37°C with shaking at 225 RPM. 1 mL of cell culture from the 50 mL conical tube was transferred into a 250 mL Erlenmeyer flask containing 100 mL of sterile LB broth with 50 µg/mL kanamycin and 34 µg/mL chloramphenicol, and this flask was incubated at 37°C with shaking at 225 RPM to an A_{600} of 0.6-0.8. Expression was then induced using 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG). Incubation at 37°C with shaking at 225 RPM was continued for 24 h. 1 mL samples were taken at 2 h, 4 h, 6 h, and 24 h post-induction. Samples were centrifuged at 5000 g (in a microcentrifuge) for 10 min, the supernatants discarded, and the pellets resuspended in 100 µL TEN buffer (20 mM Tris pH 8.0, 1 mM EDTA, 100 mM NaCl); resuspended pellets were flash frozen in liquid nitrogen and stored at -80°C.

2.2.2 Lysis and Centrifugation

Resuspended cell pellets were subjected to 5 cycles of freezing in liquid nitrogen and thawing in a 37°C water bath. 5 µL of 3 mg/mL DNase I was then added to each microcentrifuge tube, and mixed by gently inverting the tubes 30 times. After a 20 min incubation period, samples were subjected to an additional 5 freeze-thaw cycles. Soluble and insoluble fractions were separated by centrifugation at 16300 g (in a microcentrifuge) for 15 min, and the supernatants were transferred to new tubes. The pellets were resuspended in TEN buffer and all tubes were flash frozen in liquid nitrogen and stored at -80 °C.

2.2.3 SDS-PAGE and Densitometry

Table 2.1: **SDS-PAGE gel recipes.**

	Stacking Gel (5% Acrylamide)	**Resolving Gel (12% Acrylamide)**
Deionized Water	3.6 mL	4.3 mL
40% Acrylamide/ Bis-Acrylamide	625 µL	3 mL
0.5 M Tris-HCl, pH 6.8	625 µL	
1.5 M Tris-HCl, pH 8.8		2.5 mL
10% SDS	50 µL	100 µL
Ammonium Persulfate	50 µL	100 µL
Tetramethylethylenediamine	10 µL	10 µL

Stacking gels (5% acrylamide) and resolving gels (12% acrylamide) were mixed according to the recipes in Table 2.1. The resolving gel was poured first, and allowed to polymerize for 40 min before the stacking gel was poured on top. The stacking gel was allowed to polymerize for 40 min before the samples were loaded. 20 µL of each sample was mixed with 20 µL loading buffer (2% w/v SDS, 10% w/v glycerol, 0.1% bromophenol blue, 50 mM Tris-HCl, pH 6.8) and 4 µL β-mercaptoethanol, and boiled for 10 minutes. 15 µL of each mixture was loaded into the appropriate lane.

A constant voltage of 150 V was applied for approximately 2 hours (until the dye front was less than 1 cm from the bottom of the 10 cm by 8 cm gel. The gel was stained using Coomassie Blue (625 mL deionized water, 300 mL methanol, 75 mL glacial acetic acid, 1 g Coomassie Blue dye) overnight, and then destained (625 mL deionized water, 300 mL methanol, 75 mL glacial acetic acid) until the background was clear.

The destained gels were imaged using a BIS303PC gel documentation system (DNR Bio-Imaging Systems) and pixel densities were quantitated using the TotalLab 100 software package (Nonlinear Dynamics). The background (an approximation of the pixel density of protein-free gel) was determined using the "rolling ball" method and subtracted from the Adnectin™ bands. The percentage of Adnectin™ in the insoluble fraction of each sample was calculated according to Eq. 2.1, where S represents the integrated pixel density in the soluble Adnectin™ band, and I represents the integrated pixel density in the insoluble Adnectin™ band.

$$\% \; Insoluble = 100 \cdot \frac{I}{I+S} \tag{2.1}$$

2.3 Results & Discussion

N-terminal gluconoylation is a common post-translational modification of Adnectins™ [121]. The gluconoylated and non-gluconoylated forms differ slightly in molecular weight, and two Adnectin™ bands are frequently distinguishable in a single SDS-PAGE gel lane. The fraction of protein gluconoylated is approximately the same in each lane (both soluble and insoluble) (Figure 2.1), leading us to conclude that gluconoylation is not an important determinant of IB formation. Measurements of the percentages of Adnectin™ found in the soluble and insoluble fractions (Table 2.2) included both the gluconoylated and non-gluconoylated forms.

Assuming a positive charge on arginine/lysine side chains, a negative charge on aspartate/glutamate side chains, and no charge on histidine side chains, each Adnectin™ has a net charge of +1 outside of the FG loop (all of the FG loop sequences include charged residues), with the exception of 5898_C01, which has a net charge of +2 outside of the FG loop. The gluconoyl group may be phosphorylated [123, 124], modifying the net charge on the protein. The difference in molecular weight resulting from gluconoylation is small, and would not normally be expected to result in such a clear distinction between gluconoylated and non-gluconoylated bands; it is likely that phosphogluconoylation of the protein interferes with SDS binding. The analysis in Section 4.3.3 demonstrates that the net charge on the Adnectins™ is a poor predictor of IB formation, which is consistent with the observation that the fraction of protein (phospho)gluconoylated appears to be the same in soluble and insoluble lanes. Inspection of Table 2.2 reveals that the hydrophobicity of the first, second, and sixth FG loop residues is correlated with IB formation. This relationship will be systematically explored in Section 4.3.

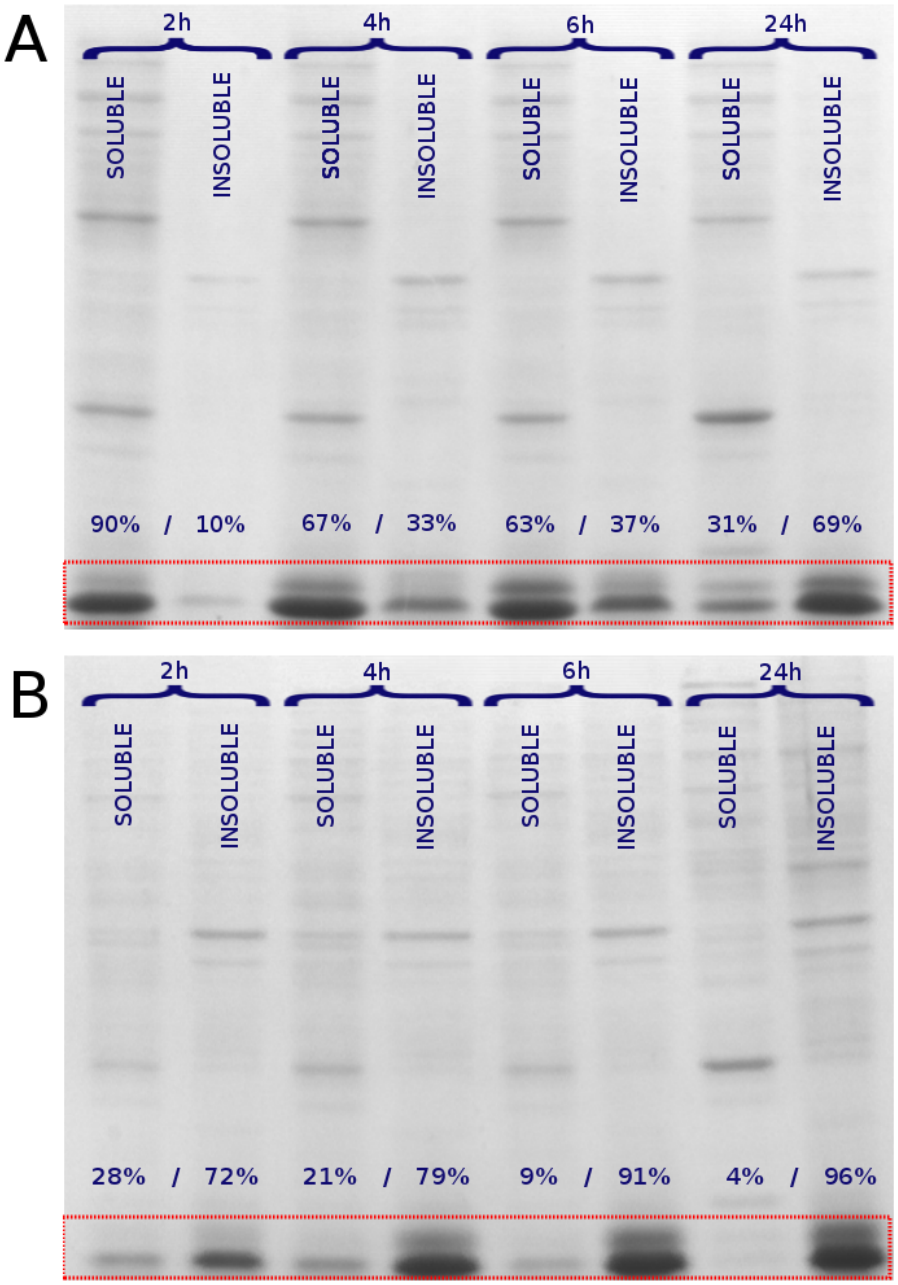

Figure 2.1: Pictures of SDS-PAGE gels representative of those analyzed by densitometry in order to produce the data in Table 2.2. A: relatively soluble mutant 6199_B02, B: relatively insoluble mutant 6199_G07. Adnectin bands are indicated by red rectangles. The gel pictures (as shown here) have been cropped, and the contrast/brightness adjusted for consistency using GIMP.

Table 2.2: **Average IB formation post-induction (two growths; average ± range).**

Adnectin™	**FG Loop Sequence**	**2h (% IB)**	**4h (% IB)**	**6h (% IB)**	**24h (% IB)**
6199_B07	ERRDYR	20 ± 2	17 ± 1	19 ± 2	57 ± 4
5898_B01	KMRDYR	17 ± 7	19 ± 3	25 ± 2	67 ± 5
6199_B01	GSRDYE	15 ± 3	20 ± 0	26 ± 6	58 ± 3
6199_E01	RSRDYR	17 ± 7	21 ± 7	23 ± 3	57 ± 2
5898_C01	SLRDYG	20 ± 1	22 ± 0	33 ± 0	83 ± 7
5898_H01	NLRDYG	26 ± 4	24 ± 7	36 ± 3	78 ± 1
6199_A03	KVRDYR	29 ± 7	25 ± 9	31 ± 5	81 ± 10
6199_D06	SRRDYG	29 ± 2	27 ± 2	27 ± 5	75 ± 0
6199_B04	CRRDYG	28 ± 6	35 ± 2	39 ± 1	62 ± 3
6199_B05	EMRDYG	22 ± 9	38 ± 2	44 ± 4	84 ± 1
6199_B02	TQRDYG	21 ± 11	40 ± 6	37 ± 0	77 ± 8
5898_E01	SLRDYA	29 ± 3	42 ± 1	56 ± 1	90 ± 3
6199_E02	HFRDYG	38 ± 3	44 ± 6	55 ± 10	80 ± 6
6199_E06	RLRDYE	36 ± 3	47 ± 2	61 ± 1	89 ± 2
5898_F01	MSRDYG	46 ± 2	47 ± 3	61 ± 3	93 ± 2
6199_D08	VLRDYR	47 ± 8	49 ± 3	55 ± 1	91 ± 5
6199_D01	RIRDYG	36 ± 4	51 ± 2	62 ± 4	85 ± 7
6199_E03	KLRDYL	47 ± 5	53 ± 1	63 ± 3	90 ± 2
6199_H04	LFRDYG	52 ± 5	54 ± 5	54 ± 7	70 ± 4
6199_F07	SLRDYV	40 ± 4	58 ± 3	64 ± 6	92 ± 3
5898_C02	LLRDYG	47 ± 2	59 ± 0	67 ± 5	88 ± 11
6199_F01	DYRDYL	46 ± 2	59 ± 1	65 ± 2	89 ± 3
6199_D07	ALRDYV	57 ± 4	60 ± 0	67 ± 4	91 ± 4
6199_H07	QLRDYS	57 ± 4	64 ± 1	76 ± 7	93 ± 3
6199_C05	LVRDYG	51 ± 7	65 ± 3	71 ± 1	nd
6199_B03	TWRDYL	61 ± 1	69 ± 3	72 ± 2	91 ± 5
6199_F08	TLRDYM	55 ± 9	70 ± 6	77 ± 7	94 ± 2
6199_A05	YLRDYT	62 ± 7	73 ± 3	83 ± 1	96 ± 1
6199_D05	FIRDYG	63 ± 8	73 ± 7	76 ± 3	91 ± 2
6199_G07	LIRDYG	66 ± 6	74 ± 5	82 ± 9	92 ± 4
6199_A07	LLRDYV	68 ± 2	76 ± 2	84 ± 3	95 ± 2

Quantification of the amounts of Adnectin™ in the soluble and insoluble fractions of each sample is fraught with many potential sources of error. All of the mutants were grown in the same incubator, using the same temperature and shaking settings (Section 2.2). The mutants were grown in groups of 5-6, and induced simultaneously; OD600 measurements at the time of induction were always between 0.6 and 0.8, but varied to some extent between mutants in the same growth group. Minor variations in the time taken to gather and process samples prior to flash-freezing were unavoidable.

Lysis and separation of the soluble/insoluble fractions by centrifugation were generally problem-free. SDS-PAGE gels were screened for signs of imperfect separation by inspection of several prominent non-Adnectin™ bands. Just one of these bands, thought to be an outer membrane protein (OMP) [125], normally appears in the insoluble lanes. The presence of soluble proteins in an insoluble lane is diagnostic of incomplete lysis, while the presence of OMP in a soluble lane is evidence of either incomplete DNA digestion (which diminishes the effectiveness of centrifugation) or inadvertent resuspension of insoluble material during post-centrifugation soluble fraction transfer.

The amount of protein in each band was inferred from gel image pixel density, relative to a protein-free background reference that varied from one lane to the next. We elected to use only relative measurements (percentage of total Adnectin™ in the soluble and insoluble lanes for each time point), which eliminates the error inherent in calibrating densitometric measurements using a band with a known amount of protein as a standard. However, the percentages that we have reported in Table 2.2 do assume a linear relationship between pixel density and amount of protein. Any error introduced by this assumption will not be revealed by repeat experiments.

Table 2.3: **Correlation coefficients (R) of IB formation at different time points.**

	24h (% IB)	**6h (% IB)**	**4h (% IB)**	**2h (% IB)**
2h (% IB)	0.739	0.925	0.944	1.000
4h (% IB)	0.787	0.976	1.000	
6h (% IB)	0.850	1.000		
24h (% IB)	1.000			

Considering the difficulty of exactly replicating the experimental conditions, the repeatability of the quantification of Adnectin™ IB formation (Table 2.2) is quite good. A handful of larger than average variations can be found in the 2 and 6 hour time point data; the 4 hour time point data are somewhat superior in this regard. The IB formation data from the 2, 4, and 6 hour time points are highly correlated (Table 2.3). The 24 hour time

point data are only moderately correlated with the other time points, which is not unexpected given the large time difference. The growth procedure is designed to induce protein expression while the number of *E. coli* cells is growing exponentially (log phase); after 24 hours, the metabolic state of the bacteria may be quite different. The average proportion of Adnectin™ found in IBs at 4 hours post-induction varies from 17% (6199_B07) to 76% (6199_A07), despite the fact that these two sequences differ only in three positions (the first, second, and sixth residues of the FG loop). In Chapters 3, 4, and 5, we will investigate the stability and aggregation propensity of the Adnectins™, in an effort to reveal the factors that determine their varying propensities to form IBs.

Chapter 3

Stability and Thermal Aggregation Analysis by Differential Scanning Calorimetry

3.1 Introduction

Polypeptides in fully or partially unfolded conformations are susceptible to aggregation (Chapter 1). One frequently observed predictor of aggregation propensity is thermodynamic stability [126, 127, 128], which determines the proportion of polypeptides in natively folded vs. denatured conformations at equilibrium. Most evolved proteins have stabilities in the range of 5-15 kcal/mol (21-63 kJ/mol) [129], implying that >99.9% of the polypeptides are natively folded at equilibrium. Under these conditions, it is improbable that two or more unfolded polypeptides will encounter each other and have an opportunity to initiate aggregation. Mutants with decreased stability may display a higher propensity to aggregate because a greater proportion of polypeptides are unfolded at equilibrium. Furthermore, if mutations that alter the stability of the native state similarly affect the transition state, a correlation between folding rate and thermodynamic stability may be observed (i.e. following synthesis, a destabilized mutant may remain fully or partially unfolded longer on average, even if no significant fraction is unfolded at equilibrium).

Thermodynamic stability can be measured by a variety of methods, including differential scanning calorimetry (DSC). DSC entails measurement of the energy required to heat a sample cell (containing protein and buffer) relative to that required to heat a reference

cell (containing only buffer) to the same temperature [130]. Protein unfolding is an endothermic process; more energy per degree Celsius is required to raise the temperature of the sample cell in the transition region (where an increase in temperature is accompanied by a measurable increase in the proportion of protein that is denatured). After appropriate baseline subtraction, the midpoint of thermal denaturation (T_m) can be identified as the temperature at which the endothermic peak reaches its maximum, and the area under the curve is directly related to the calorimetric enthalpy of unfolding (ΔH_{cal}).

At any temperature in the transition region, the sample cell contains a mixture of folded and unfolded proteins. If the unfolding can be modelled as a reversible two-state transition, an independent measure of the enthalpy of unfolding can be obtained using the van't Hoff equation (Eq. 3.1) [131]. The van't Hoff enthalpy (ΔH_{VH}) depends only on the shape of the endothermic peak (from which the equilibrium constant as a function of temperature can be determined), and thus has the advantage of being insensitive to errors in protein concentration, as well as any impurities (including unfolded or misfolded protein) that do not interfere with the two-state transition.

$$\frac{d\ ln(K_{eq})}{dT} = \frac{\Delta H_{VH}}{RT^2} \tag{3.1}$$

In addition to exploring the possibility that Adnectin™ IB formation is correlated with thermodynamic stability, we have used DSC to investigate the aggregation propensity of unfolded Adnectins™. When aggregation is observed by DSC, it often manifests as a strongly exothermic signal that appears when the concentration of thermally unfolded protein exceeds a critical threshold [130]. The overlap of endothermic protein unfolding and exothermic aggregation renders it difficult to extract thermodynamic parameters pertaining to either process from the data. However, clues regarding the aggregation propensity of unfolded Adnectins™ can be gleaned by varying the experimental conditions (pH, protein concentration), and by comparing different mutants under identical conditions.

3.2 Methods

3.2.1 Expression and Purification

Overnight cultures were prepared in 50 mL conical tubes as detailed in Chapter 2. For each Adnectin™, 1 mL of cell culture from the 50 mL conical tube was transferred into a 4 L Erlenmeyer flask containing 1 L of sterile LB broth with 50 µg/mL kanamycin and

34 µg/mL chloramphenicol, and this flask was incubated at 37°C with shaking at 225 RPM to an A_{600} of 0.6-0.8. Expression was then induced using 1 mM IPTG. After 2-4 h post-induction, cells were pelleted by centrifugation for 20 min at 5000 g and 4°C. The supernatant was poured off, and the cells were resuspended in 40 mL of a buffer consisting of 50 mM sodium phosphate, pH 8.0, 0.5 M NaCl, and 25 mM imidazole. The cells were then lysed by sonication on ice using four 15 second pulses (60 W) separated by 10 second pauses, using a W-225R probe sonicator with a standard tapered microtip attached to a 1/2" disruptor horn (Heat Systems-Ultrasonics Inc.).

The cell lysate was centrifuged for 30 min at 20000 g to pellet insoluble material. Following the addition of 500 µL of 3 mg/mL DNase I, the supernatant was incubated for 20 min at room temperature, and then syringe filtered (0.45 µm Supor® membrane, Pall Corporation). Adnectins™, which have a 6-residue C-terminal polyhistidine tag, were purified from the supernatant by nickel affinity chromatography, and then exchanged into 20 mM sodium acetate pH 4.0 buffer (except where otherwise noted) and concentrated (see Table B.1 for concentrations) using an Amicon Ultra-4 centrifugal filter (EMD Millipore) with a 3.5 kDa molecular weight cut-off. Protein concentrations were determined by absorbance at 280 nm.

3.2.2 Differential Scanning Calorimetry

DSC measurements were performed using a MicroCal VP-DSC (MicroCal LLC). Unless otherwise stated, samples were in 20 mM sodium acetate pH 4.0, flow-through from the centrifugal filter was used as the reference buffer, and the scan rate was $1°C \cdot min^{-1}$. Buffer/buffer reference scans were collected for each experiment and subtracted from protein/buffer scans. Following normalization for protein concentration, a "progress baseline" was subtracted, and the data were fit to the MN2State model using Origin 5.0 SR2 (OriginLab Corporation).

3.3 Results

Reversible unfolding is normally a prerequisite for the determination of thermodynamic parameters from DSC data, but some proteins aggregate upon thermal denaturation at physiological pH. DSC performed on an Adnectin™ with low propensity to form IBs (6199_D06; Chapter 2) at pH 7.8 and pH 6.0 resulted in irreversible aggregation (Figure 3.1). At a pH of 5.0, the outcome varied between irreversible aggregation (Figure 3.1) and reversible unfolding (Figure 3.2) in a protein concentration-dependent manner.

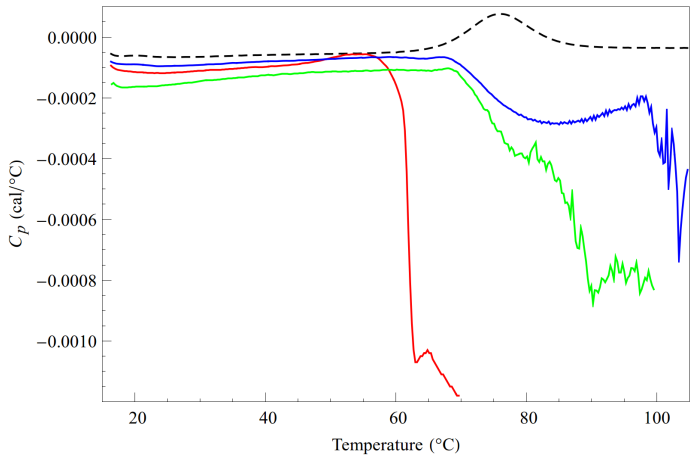

Figure 3.1: Irreversible aggregation of Adnectin™ 6199_D06 in 20 mM HEPES pH 7.8 (red), 20 mM citrate pH 6.0 (green), and 20 mM acetate pH 5.0 (blue). Reversible scan (dashed black; 20 mM acetate pH 4.0) shown for reference. Buffer/buffer reference scans subtracted. Protein concentrations are between 0.41-0.48 mg/mL (data not normalized for concentration). Figure produced using Mathematica™.

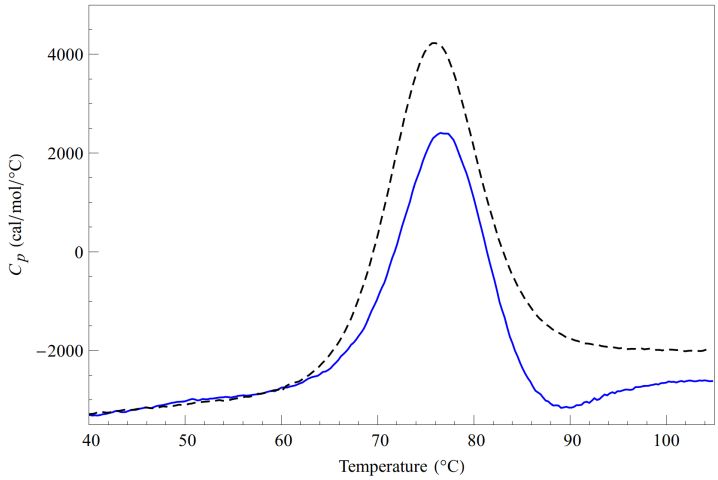

Figure 3.2: Reversible unfolding and association/dissociation of Adnectin™ 6199_D06 (0.34 mg/mL) in 20 mM acetate pH 5.0 (blue); 6199_D06 (0.41 mg/mL) in 20mM acetate pH 4.0 (dashed black) shown for reference. Buffer/buffer reference scans subtracted and data normalized for concentration. Figure produced using Mathematica™.

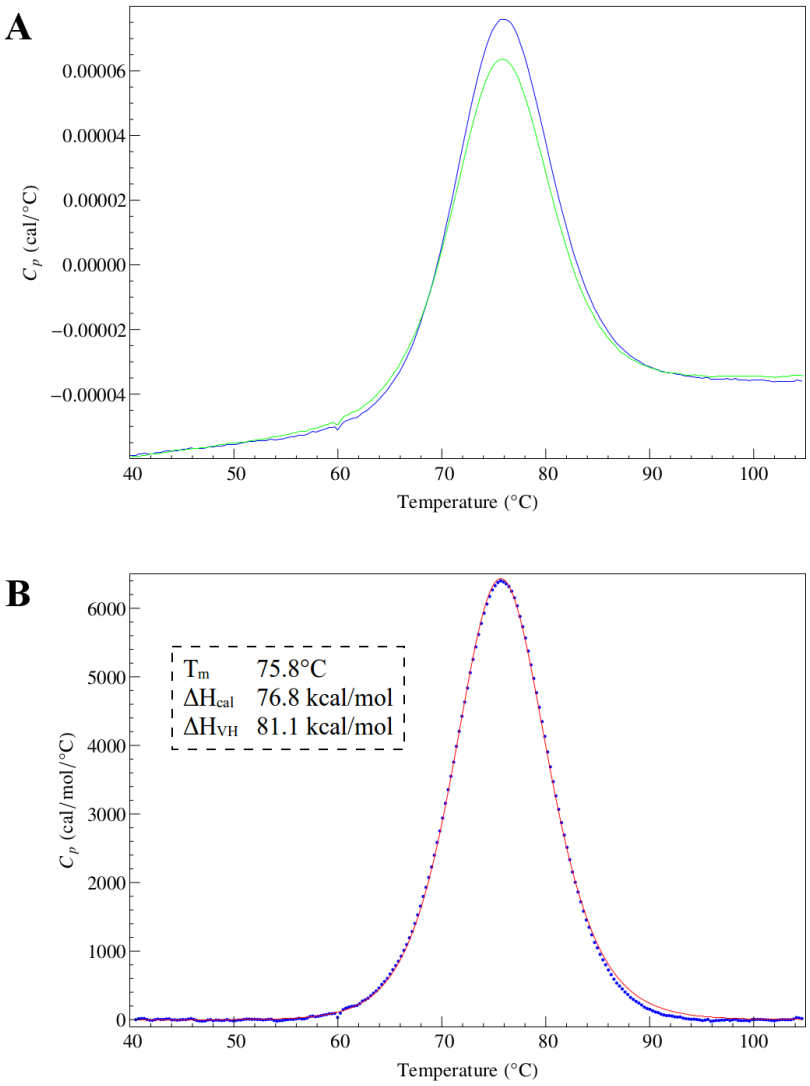

Figure 3.3: DSC data for Adnectin™ 6199_D06 (0.41 mg/mL, pH 4.0). A: Scan (blue) and rescan (green) with buffer/buffer reference scan subtracted. Not normalized for concentration. B: Normalized for concentration and baseline subtracted; data points (blue) and fit to the MN2State model (fit: red; fitted values: dashed box). Figure produced using Mathematica™.

A high degree of reversibility (verified by rescan for each Adnectin™) was observed at pH 4.0 (Figure 3.3A). The smaller endothermic peak observed upon rescan is evidence that some protein is lost to misfolding and/or degradation at elevated temperature. In order to assess the concentration dependence and repeatability of the DSC results, scans were performed at two different protein concentrations (Table 3.1, Table B.1); for the reasons laid out in Section 3.1, the comparisons between different concentrations are based on van't Hoff enthalpies rather than calorimetric enthalpies.

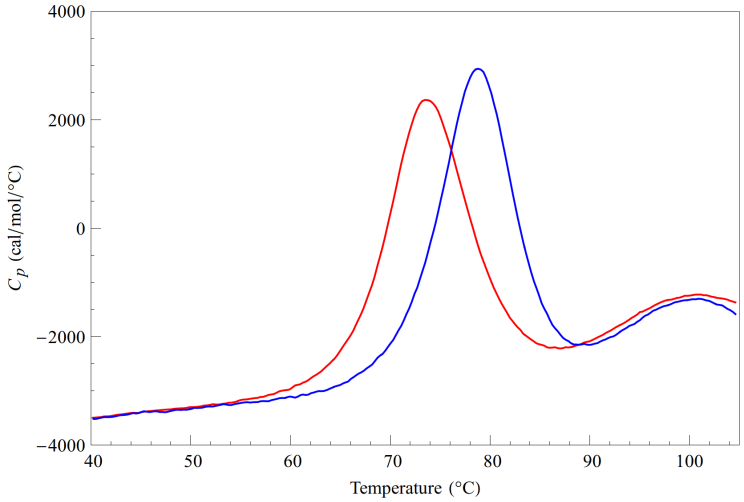

Figure 3.4: Reversible unfolding and association/dissociation of Adnectins™ 6199_G07 (red; 0.47 mg/mL) and 6199_B05 (blue; 0.43 mg/mL). Buffer/buffer reference scans subtracted and data normalized for concentration. Figure produced using Mathematica™.

Despite the high degree of reversibility observed at pH 4.0, for some Adnectins™ there is evidence of exothermic association (a word chosen to differentiate the phenomenon from irreversible aggregation) upon unfolding. The distortion of the endothermic peaks is subtle (first peaks, Figure 3.4), but is apparent in the protein concentration dependence of the T_m's (6199_G07 and 6199_B05, Tables 3.1 and B.1). Adnectins™ demonstrating this behaviour are highlighted in Table 3.1; from these data only "apparent" T_m's can be extracted, and ΔH_{VH} values cannot be accurately determined. Dissociation is observed in the form of a second endothermic peak at approximately 100°C; the T_m of dissociation does not appear to vary much, even between Adnectins™ with markedly different T_m's of unfolding (Figure 3.4). A similar phenomenon is observed in the reversible unfolding of Adnectin™ 6199_D06 at pH 5.0 (Figure 3.2), but not at pH 4.0 (Figure 3.3). The exothermic nature of the

association between unfolded 6199_D06 polypeptides at pH 5.0 is apparent from the steep drop below the expected post-transition baseline (Figure 3.2).

Table 3.1: **Summary of differential scanning calorimetry data at pH 4.0; apparent T_m's highlighted. All data are averages of two trials ± range.**

Adnectin™	**FG Loop Sequence**	**% IB 4 h**	**T_m (°C)**	**ΔH_{VH} (kcal/mol)**
6199_B07	ERRDYR	17 ± 1	78.2 ± 0.1	82.0 ± 0.9
5898_B01	KMRDYR	19 ± 3	73.0 ± 0.2	76.1 ± 0.7
6199_B01	GSRDYE	20 ± 0	81.0 ± 0.0	81.6 ± 1.2
5898_H01	NLRDYG	24 ± 7	78.5 ± 0.1	85.4 ± 3.4
6199_A03	KVRDYR	25 ± 9	72.4 ± 0.0	77.0 ± 0.9
6199_D06	SRRDYG	27 ± 2	75.8 ± 0.0	80.2 ± 0.9
6199_B05	EMRDYG	38 ± 2	79.3 ± 0.9	
6199_B02	TQRDYG	40 ± 6	78.2 ± 0.1	82.2 ± 0.5
5898_E01	SLRDYA	42 ± 1	78.8 ± 0.1	84.0 ± 0.1
6199_E06	RLRDYE	47 ± 2	74.3 ± 0.1	78.1 ± 1.4
5898_F01	MSRDYG	47 ± 3	77.6 ± 0.1	82.1 ± 0.4
6199_D01	RIRDYG	51 ± 2	73.1 ± 0.1	79.0 ± 0.6
6199_D08	VLRDYR	49 ± 3	75.4 ± 0.0	78.6 ± 0.9
6199_E03	KLRDYL	53 ± 1	74.4 ± 0.0	78.4 ± 1.6
5898_C02	LLRDYG	59 ± 0	76.8 ± 0.0	80.9 ± 2.4
6199_F01	DYRDYL	59 ± 1	82.1 ± 0.2	85.6 ± 0.4
6199_D07	ALRDYV	60 ± 0	79.1 ± 0.0	81.5 ± 1.1
6199_C05	LVRDYG	65 ± 3	75.2 ± 0.5	
6199_B03	TWRDYL	69 ± 3	78.2 ± 0.2	81.3 ± 0.4
6199_F08	TLRDYM	70 ± 6	78.6 ± 0.0	85.4 ± 2.3
6199_A05	YLRDYT	73 ± 3	76.7 ± 0.4	
6199_G07	LIRDYG	74 ± 5	72.5 ± 0.8	
6199_A07	LLRDYV	76 ± 2	75.5 ± 0.0	

3.4 Discussion

In a pH range of 7.4-7.8, similar to the intracellular pH maintained by *E. coli* [132], all Adnectins™ that we have investigated aggregate irreversibly upon thermal denaturation. Adnectin™ 6199_D06 switches from irreversible aggregation upon unfolding at pH 6.0 to highly reversible unfolding at pH 5.0 (concentration dependent; at higher concentrations irreversible aggregation may still be observed). Given the pH range, a likely cause of this shift in behaviour is the protonation of one or more histidine residues, which would discourage intermolecular association through an increase in net charge per polypeptide. The only histidine residues in 6199_D06 are the six in the C-terminal polyhistidine tag. Similarly, none of the other Adnectins™ studied by DSC have any histidine residues outside of the polyhistidine tag.

The stability of the wild-type ^{10}Fn3 domain is known to be pH-dependent as a result of the negatively charged patch formed by residues D7, E9, and D23 (with a midpoint of transition at approximately pH 4.0) [133]. Mutation of D7 to asparagine or lysine has been shown to completely eliminate this pH dependence (by lowering the perturbed E9/D23 pKa values). The corresponding residue numbers in the Adnectin™ sequences (Appendix A) are 9, 11, and 25. All of the Adnectins™ under investigation have in common a D25S mutation (relative to the corresponding wild-type amino acid; a different position within the same negatively charged patch). Although pH-dependent association and aggregation complicate measurement of Adnectin™ pH/stability curves by DSC, the apparent T_m's in Figure 3.2 are very similar; qualitatively, it appears that the D25S mutation serves to preclude the pH-dependence of stability observed in wild-type ^{10}Fn3.

All of the Adnectins™ that we have studied by DSC have proven stable enough that no significant fraction is expected to be denatured at equilibrium under physiological conditions. The lack of significant correlation between T_m and IB formation (Chapter 2) illustrated in Figure 3.5 supports this hypothesis.

Although all Adnectins™ aggregate when thermally denatured at physiological pH, different propensities to associate can be distinguished at reduced pH. Adnectins™ with high propensity to associate are more likely to aggregate when overexpressed in *E. coli*, but there are counter-examples: 6199_B05 associates when thermally denatured at pH 4.0, but has relatively low propensity to form IBs; 6199_B03 and 6199_F08 do not associate when thermally denatured at pH 4.0, despite a relatively high propensity to form IBs (Table 3.1). The intrinsic association/aggregation propensity of unfolded polypeptides is only one piece of the IB formation puzzle, but it is clearly important; we will continue to investigate the intrinsic aggregation propensity of Adnectin™ sequences in Chapter 4.

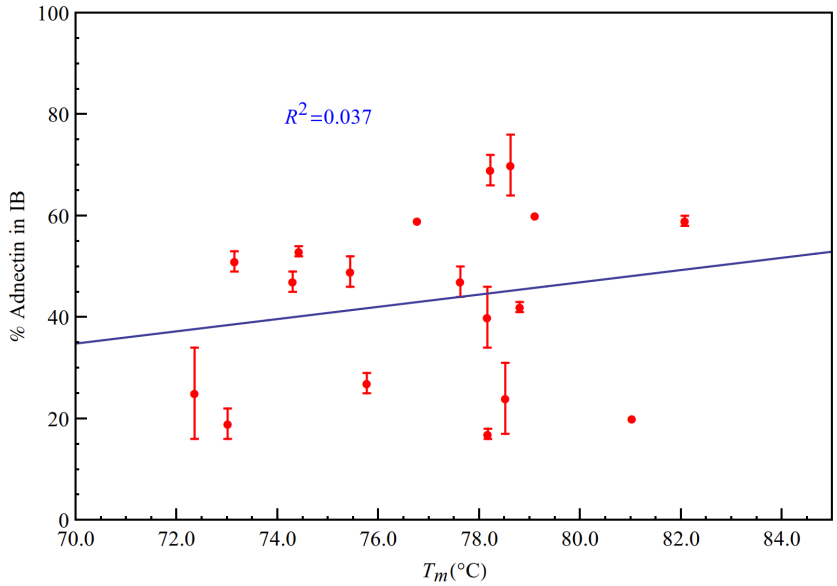

Figure 3.5: Adnectin™ IB formation data vs. midpoints of thermal denaturation. Adnectins™ highlighted in Table 3.1 excluded. Figure produced using Mathematica™.

Chapter 4

Sequence-Based Prediction of Inclusion Body Formation

4.1 Introduction

In this chapter, we will review aggregation predictions generated by the application of various sequence-based methods to the Adnectins™ (Appendix A), and compare the results with experimentally-determined IB formation propensities (Chapter 2). In undertaking this work, we had three goals:

1. To determine which method most successfully predicts Adnectin™ IB formation (potentially for use as a proxy for exhaustive experimental determination of soluble expression when screening libraries of candidate sequences, or in the making of rational protein design decisions).

2. To gain insight into the relative importance of the various physicochemical properties known to influence IB formation.

3. To generate aggregation propensity sequence profiles that can be used to help predict which specific sequence segments are most likely to participate in intermolecular interactions. These profiles will be compared to a model of sequence segment exposure based upon studies of ^{10}Fn3 (wild-type) ϕ-value analysis and mechanical unfolding.

Some of the methods considered in this chapter were designed to predict and/or validated against data specific to a certain type of aggregation (e.g. *in vitro* aggregation or

amyloid formation), but different forms of protein aggregation are likely to have many predictors in common. Furthermore, the success or failure of aggregation type-specific methods to predict Adnectin™ IB formation can be a source of insight into the formation and structure of Adnectin™ IBs.

The methods also span several different levels of abstraction, from whole-sequence, almost amino acid order-independent prediction of aggregation rates (the Chiti-Dobson equation, Section 4.3), to whole-sequence, amino acid order-dependent prediction of soluble expression (Section 4.2), to sequence segment (5-7 residues) specific predictions of aggregation potential (the sliding window methods, Section 4.4).

Implicit in the methods that assign a single score to the whole sequence (Sections 4.2 and 4.3) is the assumption that all of the segments are equally exposed and available to aggregate; this is not necessarily true of polypeptides involved in IB formation (see Chapter 1). To evaluate the likelihood that differences in the exposure of sequence segments play a role in determining Adnectin™ IB formation, we will map the aggregation propensity profiles created using the sliding window methods onto a model of sequence segment exposure (Section 4.4.3).

4.2 Soluble Expression Prediction Methods

Most modern methods designed for sequence-based prediction of soluble expression can trace their roots to the Wilkinson-Harris model, which evaluated the likelihood of soluble expression based upon average charge, turn-forming residue fraction, cysteine fraction, proline fraction, hydrophilicity, and total number of residues [134]. In addition to these, and other relatively unambiguous sequence features, modern methods also incorporate predicted features such as secondary structure, and number of domains [26, 135]. When these features are extracted from a set of training sequences for which soluble expression has been measured experimentally, machine-learning algorithms can be used to train classification systems to distinguish between soluble and insoluble (when expressed in *E. coli*) protein sequences.

4.2.1 Methods

SOLpro

The SOLpro solubility prediction algorithm was implemented using a two-stage support vector machine (SVM), a machine-learning architecture that non-linearly maps input vec-

tors to a high-dimensional feature space in which a "decision surface" can be constructed to support classification [136]. The SOLpro SVM was trained on data from a balanced (representing a wide range) and non-redundant set of proteins expressed in *E. coli* [135].

The Adnectin™ sequences were submitted to the SCRATCH Protein Predictor (http://scratch.proteomics.ics.uci.edu) for SOLpro analysis.

PROSO II

The PROSO II classification algorithm is organized into two layers. The first layer consists of a classifier based upon "k-mer" (monopeptide and dipeptide) frequencies, and a Parzen window-based evaluation of target sequence similarity to sequences in the soluble and insoluble data sets. The outputs of the first layer are then aggregated by the second layer [26].

The Adnectin™ sequences were submitted to the PROSO II evaluator (http://mips.helmholtz-muenchen.de/prosoII/prosoII.seam) for analysis.

4.2.2 Results

SOLpro

The SOLpro output is meant to be interpreted as an estimate of the probability of soluble expression in *E. coli* [135]. The percentage of Adnectins™ found in IBs 4 hours post-induction has been plotted against the SOLpro probabilities in Figure 4.1. No significant correlation is observed.

PROSO II

The PROSO II output is a solubility score ranging from 0-1 (the default threshold above which a sequence is considered likely to be soluble is 0.6) [26]. The percentage of Adnectins™ found in IBs 4 hours post-induction has been plotted against the PROSO II scores in Figure 4.2. No significant correlation is observed.

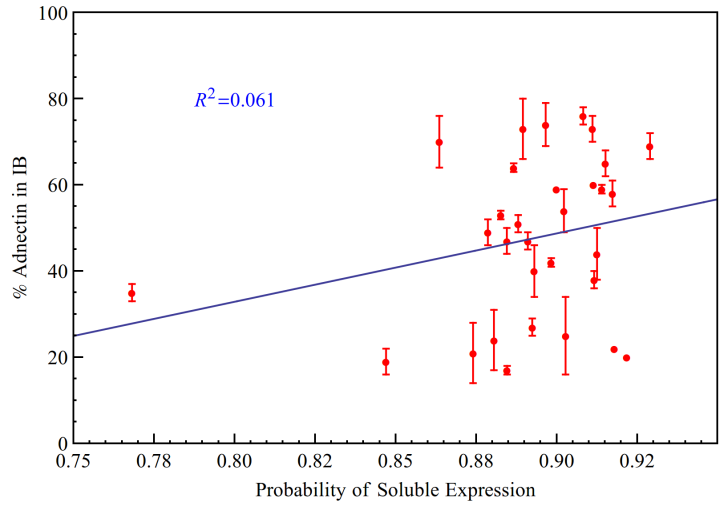

Figure 4.1: Percentage of Adnectins™ found in IBs at 4 hours post-induction (average of two growths; error bars indicate range) vs. SOLpro predicted probability of soluble expression. Line of best fit shown in blue. Figure produced using Mathematica™.

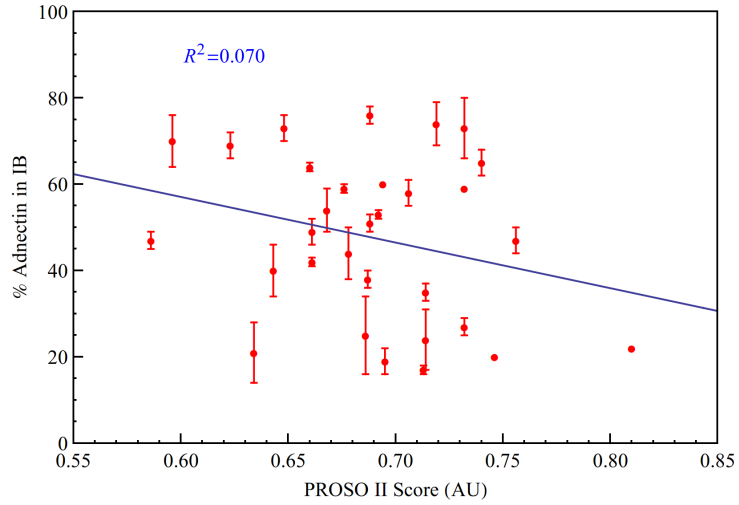

Figure 4.2: Percentage of Adnectins™ found in IBs at 4 hours post-induction (average of two growths; error bars indicate range) vs. PROSO II solubility scores. Line of best fit shown in blue. Figure produced using Mathematica™.

4.2.3 Discussion

SOLpro and PROSO II were specifically designed to predict the solubility of recombinant proteins expressed in *E. coli*, but they both fail to discriminate between Adnectins™ with substantially different soluble/insoluble expression profiles. Though the implementation details differ, both methods were trained using large sets of sequences that were carefully curated to ensure that diverse proteins with non-redundant sequences were included. Our results indicate that methods trained in this way are unlikely to be capable of accurately predicting differences in the soluble expression of closely related mutants.

4.3 Chiti-Dobson Equation

The Chiti-Dobson equation is of the form:

$$ln(v_{mut}/v_{wt}) = A\Delta Hydr. + B(\Delta\Delta G_{coil-\alpha} + \Delta\Delta G_{\beta-coil}) + C\Delta Charge \qquad (4.1)$$

It describes a logarithmic relationship between the aggregation rate of a mutant protein (v_{mut}) relative to that of a "wild-type" reference (v_{wt}), and the weighted sum of three parameters: the change in hydrophobicity ($\Delta Hydr.$), the change in propensity to switch from α-helical to β-sheet secondary structure ($\Delta\Delta G_{coil-\alpha} + \Delta\Delta G_{\beta-coil}$), and the change in net charge ($\Delta Charge$). The equation was inspired by the observation that although all three physicochemical parameters included in Eq. 4.1 correlate significantly with the $ln(v_{mut}/v_{wt})$ observed upon mutation of human muscle acylphosphatase (AcP), no single parameter offered sufficient explanatory value when considered alone. In the original publication [14], the slopes of linear models fit to plots of experimentally measured AcP $ln(v_{mut}/v_{wt})$ vs. the individual parameters were taken as the coefficients of Eq. 4.1 (Table 4.1). These coefficients were recalculated (using the same methodology) in a later study that included a wider range of experimental data [137] (Table 4.1).

Table 4.1: **Chiti-Dobson equation coefficients.**

A	***B***	***C***	**Reference**
0.633	0.198	-0.491	Chiti et. al, 2003 [14]
0.95	0.18	-0.78	Wang et. al, 2008 [137]

4.3.1 Methods

Amino acid hydrophobicities at neutral pH were taken from the supplementary information of [14] (based upon the amino acid water/octanol partition coefficients described in [138]). β-sheet propensities were also taken from the supplementary information of [14] (normalized from [139]). Charges were calculated on the assumption that arginine and lysine carry a positive charge, aspartate and glutamate carry a negative charge, and histidine is neutral. The predicted change of free energy difference between the α-helix and random coil ($\Delta\Delta G_{coil-\alpha}$) was calculated using the following equation:

$$\Delta\Delta G_{coil-\alpha} = RT \cdot ln(P_{\alpha}^{wt}/P_{\alpha}^{mut}) \tag{4.2}$$

Predicted α-helical propensities of the wild-type (P_{α}^{wt}) and mutant (P_{α}^{mut}) sequences were calculated using the AGADIR algorithm [140] (http://agadir.crg.es) using the following options: no N or C terminal protection, no parameter screening, pH 7, temperature 310 K, and ionic strength 0.1 M.

Adnectin™ 5898_B01 was designated the "wild-type", and $ln(v_{mut}/v_{wt})$ was calculated for every other mutant according to Eq. 4.1 (using the coefficients from Table 4.1).

4.3.2 Results

A plot of the percentage of Adnectins™ found in IBs 4 hours after induction vs. $ln(v_{mut}/v_{wt})$ values calculated using the coefficients from the original publication [14] is shown in Figure 4.3. A similar plot with values calculated using the coefficients from [137] is shown in Figure 4.4. Minor differences arise from the use of the two different sets of coefficients, but in both cases the correlation is clearly significant, with coefficients of determination (R^2) of approximately 0.70 and 0.66, respectively.

4.3.3 Discussion

The Chiti-Dobson equation is the result of a systematic attempt to reduce aggregation propensity (specifically, the difference between the natural logarithms of the aggregation rates of two closely related proteins) to a linear combination of three quantifiable physicochemical parameters. The remarkable success of this straightforward approach is difficult to overstate, considering the complexity of the phenomenon that it is intended to capture.

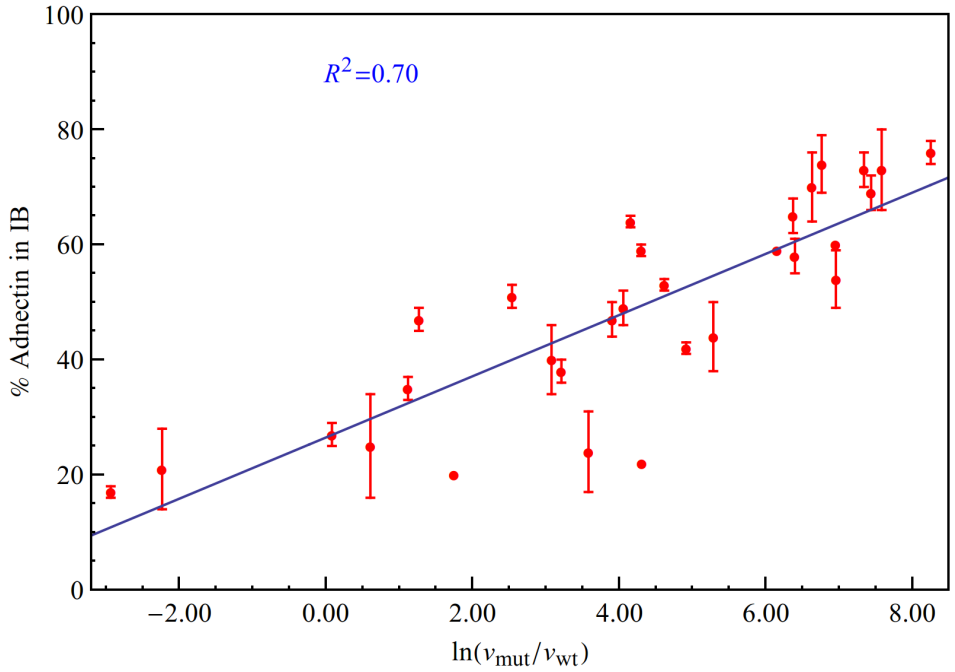

Figure 4.3: Percentage of Adnectins™ found in IBs at 4 hours post-induction (average of two growths; error bars indicate range) vs. $ln(v_{mut}/v_{wt})$ calculated using the Chiti-Dobson equation (Eq. 4.1) and the original coefficients (Table 4.1). Line of best fit shown in blue. Figure produced using Mathematica™.

Though the form of the Chiti-Dobson equation is reminiscent of the Wilkinson-Harris model (and the methods that it subsequently inspired), a key difference is the relative nature of the prediction; Eq. 4.1 was explicitly formulated to compare closely related proteins. This feature is likely the key to its success (especially relative to the methods considered in Section 4.2) in the prediction of Adnectin™ IB formation.

Table 4.2: **Chiti-Dobson parameter breakdown: correlations with Adnectin™ IB formation 4 h post-induction (independent of coefficients).**

	ΔHydr.	**$\Delta\Delta G_{coil-\alpha} + \Delta\Delta G_{\beta-coil}$**	**ΔCharge**
R^2	0.615	0.348	0.117

For our purposes, the most relevant difference between the two sets of coefficients listed in Table 4.1 is a change in the relative weighting of the secondary structure propensity

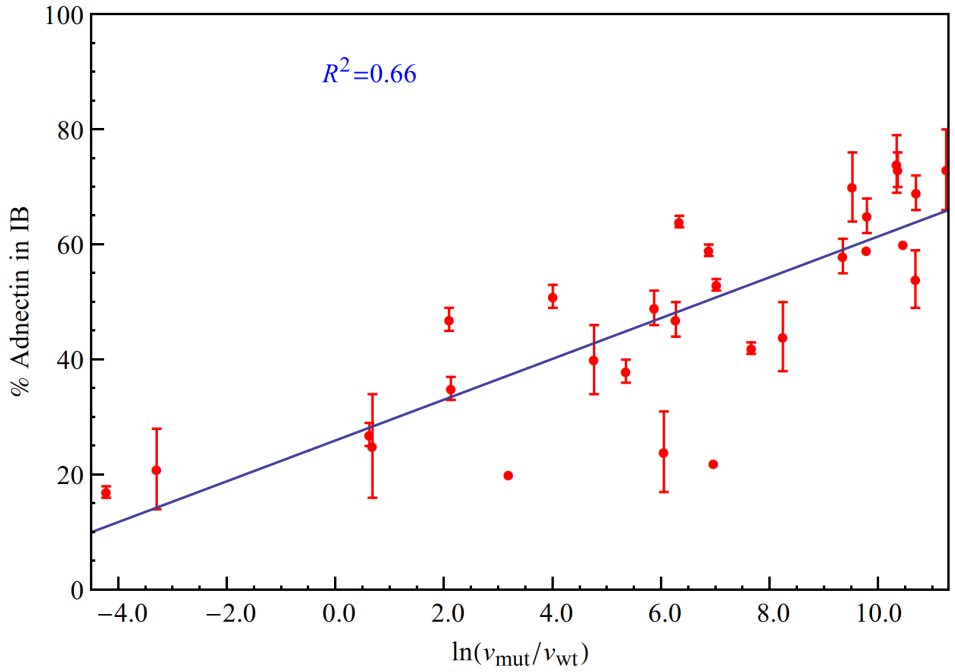

Figure 4.4: Percentage of Adnectins™ found in IBs at 4 hours post-induction (average of two growths; error bars indicate range) vs. $ln(v_{mut}/v_{wt})$ calculated using the Chiti-Dobson equation (Eq. 4.1) and the Wang-Agar coefficients (Table 4.1). Line of best fit shown in blue. Figure produced using Mathematica™.

parameter; the ratio of the hydrophobicity (A) and net charge (C) coefficients remains almost the same. The impact of the change in the relative weighting of the secondary structure propensity parameter on the correlation of the calculated $ln(v_{mut}/v_{wt})$ values with the Adnectin™ IB formation data is modest (Figures 4.3, 4.4).

The coefficients of determination that result from measuring the correlation between each of the individual Chiti-Dobson equation parameters and the percentages of Adnectins™ found in IBs 4 hours post-induction are listed in Table 4.2. It is clear that the most predictive parameter is the change in hydrophobicity, though contributions from the change in secondary structure propensity and change in net charge improve the overall correlation (Figures 4.3, 4.4). It is noteworthy that Adnectin™ IB formation is positively correlated with the intrinsic β-sheet propensity of the sequences, despite the fact that the Adnectin™ sequences must all be heavily biased towards β-sheet secondary structure as a consequence of their β-sandwich native fold. Intermolecular β-sheet secondary structure

may be an important component of Adnectin™ IBs.

4.4 Sliding-Window Methods

In recognition of the importance of relatively short (up to 7 amino acids in length) aggregation-prone segments (see Chapter 1), and because they are designed to operate on amino acid sequences without regard to higher-order structure, many useful methods rely on a "sliding window" approach that evaluates the aggregation potential of all possible segments of a given length.

The application of a group of such methods (Table 4.3) to the Adnectin™ sequences (Appendix A) has produced a data set that we have scrutinized for inter-Adnectin™ differences that correlate with IB formation (Chapter 2). We have also generated sequence position-specific aggregation propensity profiles for comparison to a model of sequence segment exposure based upon studies of ^{10}Fn3 (wild-type) ϕ-value analysis and mechanical unfolding (Section 1.2).

4.4.1 Methods

3D Profile (ZipperDB)

The task of finding sequences compatible with a known structure is sometimes approached by comparing potential sequences to a 3D profile of the structure [148, 149]. Soon after high-resolution structures of the cross-β spine of amyloid fibrils became available, this technique was adapted to determine the amyloidogenic potential of hexapeptides on the basis of how compatible their sequences are with the cross-β spine structure [141]. A 3D profile consisting of an ensemble of near-native templates [150] was derived from the GNNQQNY crystal structure. The sequence of each hexapeptide sequence segment to be evaluated is threaded into all templates and assigned a score by the Rosetta energy function [151]. The Adnectin™ sequences (Appendix A) were submitted to the ZipperDB server (`http://services.mbi.ucla.edu/zipperdb`) for analysis.

TANGO

The TANGO algorithm attempts to identify the β-aggregating regions of a protein sequence using an approach rooted in statistical mechanics [142]. For each residue, the

Table 4.3: **Sliding window aggregation prediction methods and the properties they use to determine aggregation propensity.**

Method	Basis for Prediction	Ref.
3D Profile (ZipperDB)	Compatibility of the sequence segment with the structure of the GNNQQNY crystal structure.	[141]
TANGO	A statistical mechanics algorithm. Energy calculation terms include hydrophobicity, solvation, electrostatics, and hydrogen bonding.	[142]
Waltz	Position-specific scoring matrix score, properties of amino acids, and compatibility with the GNNQQNY crystal structure.	[143]
PASTA 2.0	Hydrogen bonding statistics, and secondary structure/intrinsic disorder predictors.	[144]
AGGRESCAN	A sliding window average of amino acid aggregation propensities determined using mutational analysis.	[145]
Zyggregator	Amino acid aggregation propensity scores assigned on the basis of hydrophobicity, charge, and secondary structure formation propensity and averaged over a seven residue sliding window.	[146]
FoldAmyloid	Sliding window average (over a number of amino acids) of expected packing density and hydrogen bond formation.	[147]

occupancies of each of four conformational states (including intermolecular β-sheet aggregate) are calculated according to the Boltzmann distribution. Terms accounting for hydrophobicity, solvation, electrostatics, and hydrogen bonding are included in the energy calculations. The Adnectin™ sequences (Appendix A) were analyzed using TANGO version 2.3.1 (downloaded from http://tango.crg.es), with the following options selected: no terminal protection, pH 7.4, temperature 310 K.

Waltz

Waltz is a web-based tool that determines the amyloidogenic potential of hexapeptides using the combination of a position-specific scoring matrix-based (PSSM) sequence score, a physical properties (of the amino acids) term, and a structural modelling term derived from analysis of the GNNQQNY fibril crystal structure [143]. Waltz examines hexapeptide segments for amyloidogenicity specifically; it attempts to distinguish fibril-forming hexapeptides from those that form aggregates with an amorphous appearance or that do not aggregate at all. The Adnectin™ sequences (Appendix A) were submitted to the Waltz server (http://tango.crg.es) for analysis, with the following options selected: 90% specificity threshold, and pH 7.0.

PASTA 2.0

The PASTA 2.0 algorithm uses a pairwise energy potential derived from β-strand hydrogen bonding statistics, complemented by intrinsic disorder and secondary structure predictors, to evaluate the energy of all possible cross-β sequence segment pairings, both parallel and anti-parallel. The Adnectin™ sequences (Appendix A) were submitted to the PASTA 2.0 server (http://protein.bio.unipd.it/pasta2) for analysis, with a 90% specificity threshold specified.

AGGRESCAN

Sequence segments that are known to be aggregation-prone are sometimes described as "hot spots", and the relative aggregation propensities of individual amino acids have been characterized through mutational analysis of a model hot spot in Aβ42 [152]. The AGGRESCAN server [145] (http://bioinf.uab.es/aggrescan) was used to calculate a sliding window average (seven residues wide) of these experimentally determined amino acid aggregation propensities for each Adnectin™ (Appendix A).

Zyggregator

Zyggregator calculates the intrinsic aggregation propensity of each amino acid in a protein as a linear combination of hydrophobicity, charge, and secondary structure formation propensity scores [146, 153]. These per-residue aggregation propensities are then averaged over a sliding window of seven residues, and combined with additional terms accounting for patterns of alternating hydrophobic and hydrophilic residues, and the presence of charged "gatekeeper" residues. The Adnectin™ sequences (Appendix A) were submitted to the Zyggregator server (http://www-vendruscolo.ch.cam.ac.uk/ZaggZtox.php) with a pH of 7.0 selected.

FoldAmyloid

Strong hydrogen bonds formed between densely packed β-strands are known to be at the heart of the cross-β spine [60]. FoldAmyloid draws on amino acid packing density and hydrogen bond formation statistics captured from a database of protein structures to predict the amyloidogenicity of protein regions solely on the basis of primary sequence [147]. The Adnectin™ sequences (Appendix A) were submitted to the FoldAmyloid server (http://bioinfo.protres.ru/fold-amyloid) for analysis, with the following options selected: Scale, expected number of contacts 8 Å; averaging frame, 5; threshold, 21.4.

4.4.2 Results

The Adnectin™ sequences differ from each other only in the six residues of the FG loop, except for 5898_C01 and 5898_F01, which also differ in the 4 residues of the DE loop. Because of this high degree of sequence identity, the aggregation propensities determined by the sliding window methods differ predominantly in a narrow region centered on the FG loop (residues 79-84).

For each method, a figure mapping the aggregation propensity of representative Adnectins onto their sequences is presented below. Also, IB formation data from Chapter 2 is plotted against a measure of the overall aggregation propensity (an average, or equivalent, of the per-residue or per-segment scores), in order to gauge how well each method can discriminate between the Adnectin™ sequences.

3D Profile Method (ZipperDB)

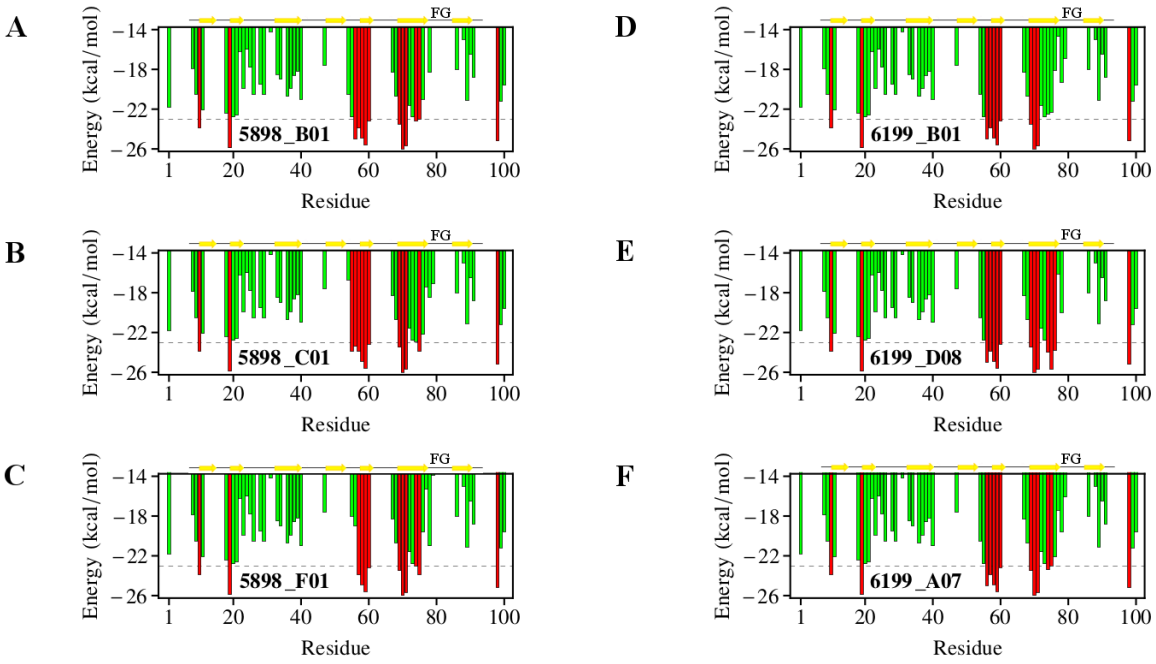

Figure 4.5: Adnectin™ sequence segment aggregation propensity data produced using the 3D Profile method. Energies below -23.0 kcal/mol (dashed line) are considered amyloidogenic. A: 5898_B01 (FG loop sequence "KMRDYR"), B: 5898_C01 FG loop sequence "SLRDYG") , C: 5898_F01 (FG loop sequence "MSRDYG"), D: 6199_B01 (FG loop sequence "GSRDYE"), E: 6199_D08 (FG loop sequence "VLRDYR"), F: 6199_A07 (FG loop sequence "LLRDYV"). Figure produced using Mathematica™.

The energy scores in Figure 4.5 are a measure of the compatibility of the various sequence segments with the crystal structure of amyloid formed by the GNNQQNY peptide. A lower score indicates higher compatibility with the GNNQNNY structure (and thus, potentially greater amyloidogenicity). The line of best fit drawn in Figure 4.6 slopes in the expected direction (with lower average scores corresponding to higher IB formation), but the scatter of the data points relative to the line is substantial.

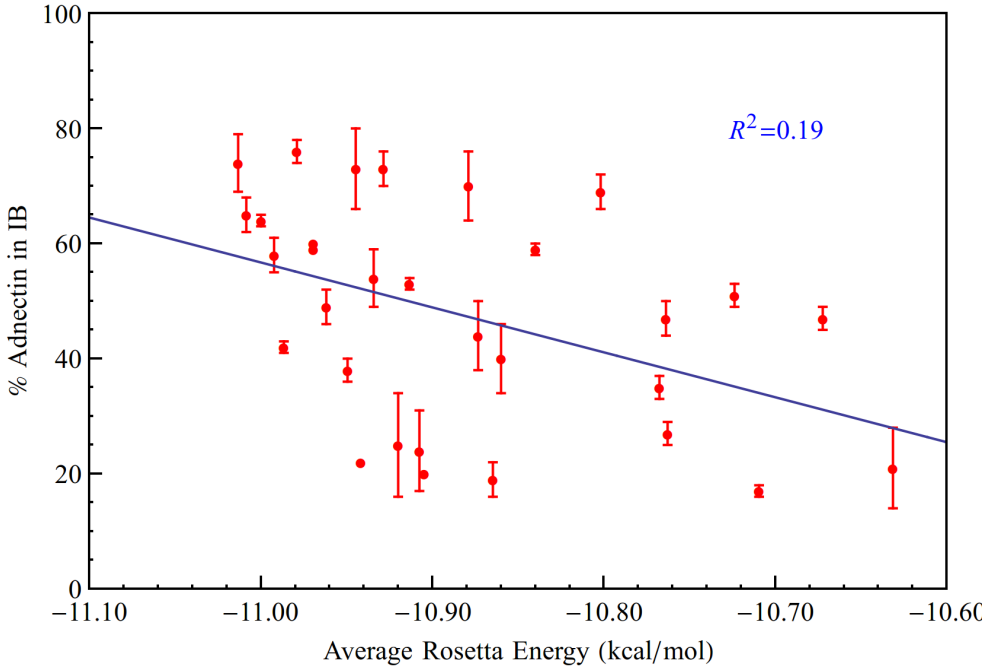

Figure 4.6: % Adnectin™ in IBs at 4 hours post-induction (average of two growths; error bars indicate range) vs. average segment scores generated by the 3D Profile method. Blue line: linear fit of the data. Figure produced using Mathematica™.

TANGO

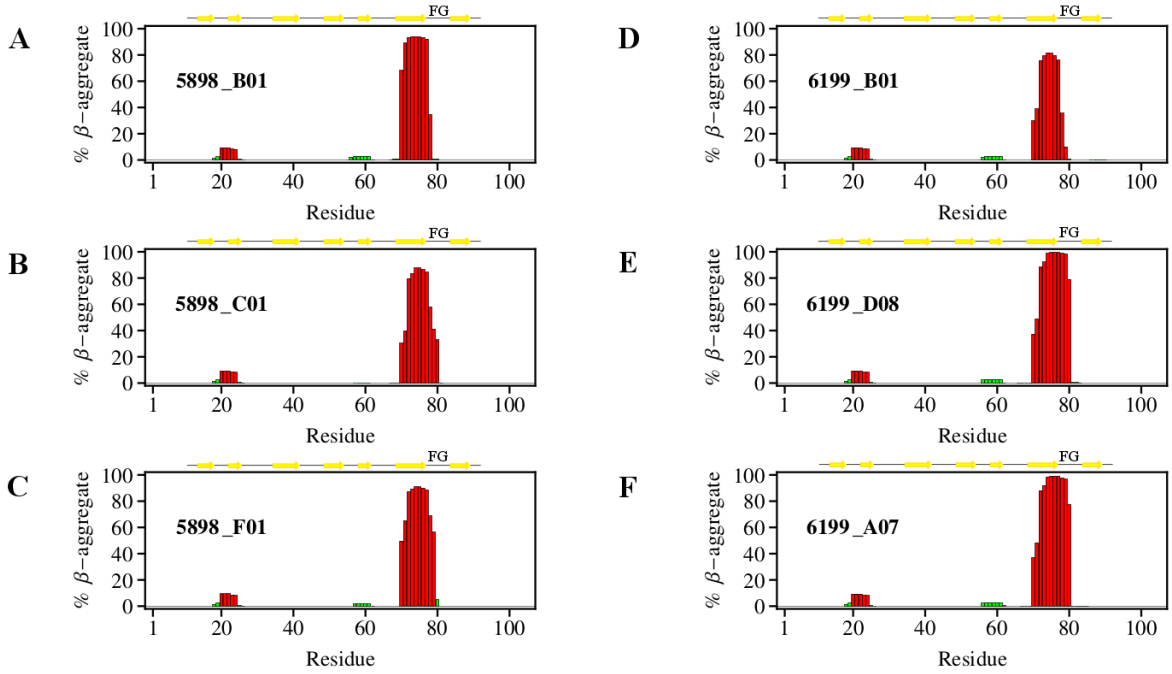

Figure 4.7: Adnectin™ sequence segment aggregation propensity profile produced using TANGO. Regions with five consecutive residues with TANGO scores greater than 5% (red) are considered aggregation-prone. A: 5898_B01 (FG loop sequence "KMRDYR"), B: 5898_C01 FG loop sequence "SLRDYG") , C: 5898_F01 (FG loop sequence "MSRDYG"), D: 6199_B01 (FG loop sequence "GSRDYE"), E: 6199_D08 (FG loop sequence "VLRDYR"), F: 6199_A07 (FG loop sequence "LLRDYV"). Figure produced using Mathematica™.

For each sequence segment, TANGO calculates the percentage occupancy of the β-aggregate state using a multiple sequence approximation. Five consecutive residues predicted to have greater than 5% occupancy of the β-aggregate state are considered an aggregation-prone segment. It is clear from Figure 4.7 that the region including β-strand F and the first residues of the FG loop has the highest predicted aggregation propensity. Figure 4.8 shows the correlation of IB formation (Chapter 2) with the overall TANGO scores. The upward slope of the line of best fit is indicative of a positive correlation between higher average occupancy of the β-aggregate state (predicted by TANGO) and IB formation, but many points fall far from the line.

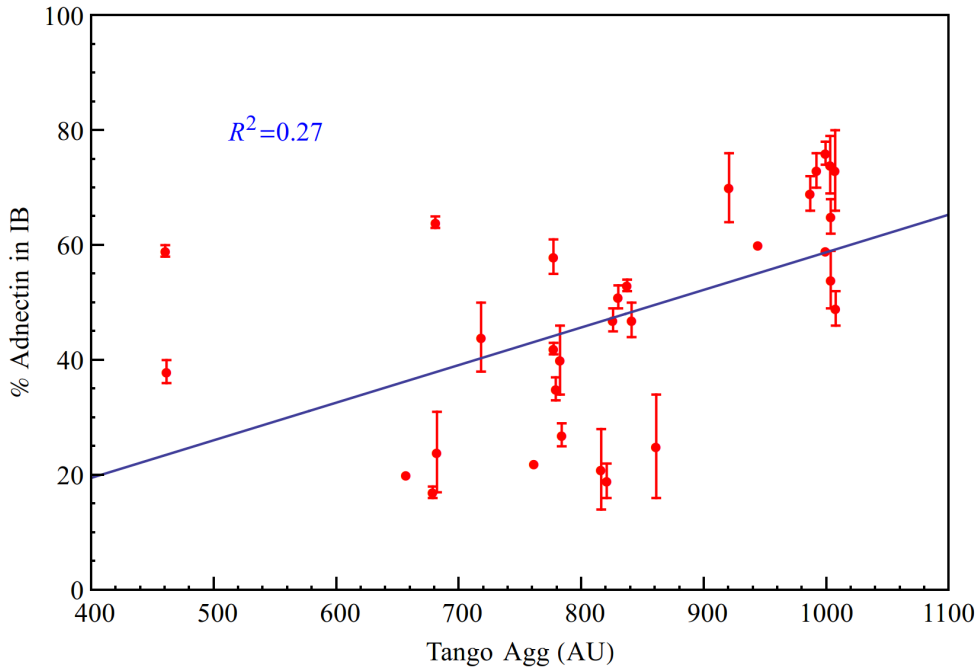

Figure 4.8: % Adnectin™ in IBs at 4 hours post-induction (average of two growths; error bars indicate range) vs. the sum of per-residue percentage β-aggregate state occupancy segment scores generated using TANGO. Blue line: linear fit of the data. Figure produced using Mathematica™.

Waltz

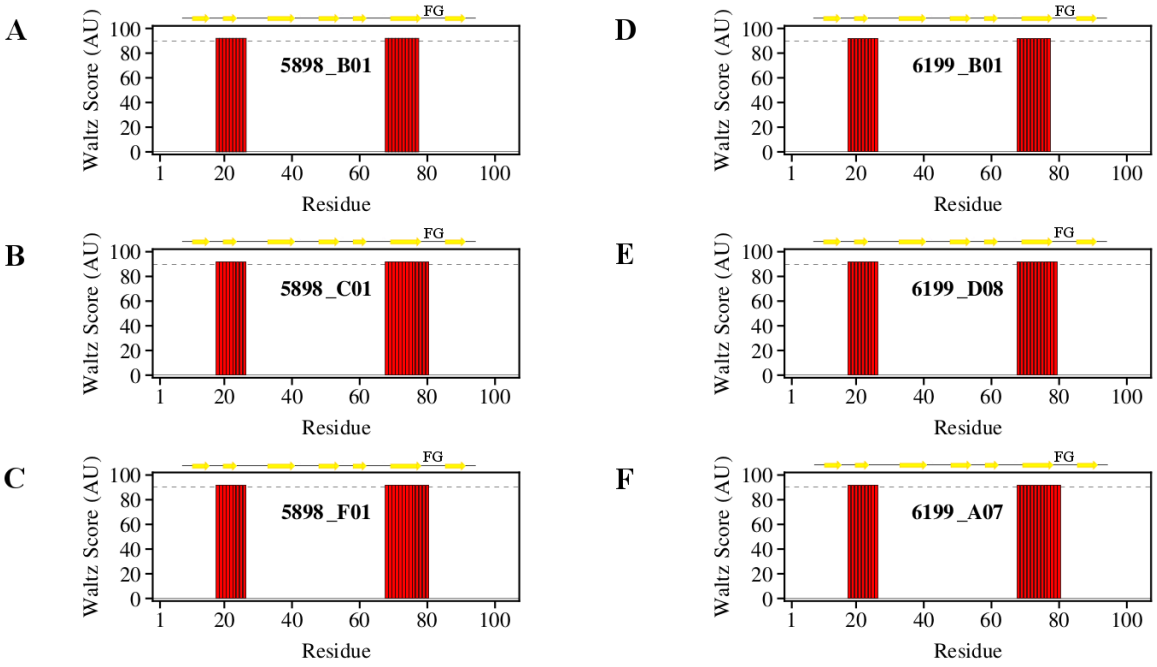

Figure 4.9: Adnectin™ sequence segment aggregation propensity profile produced using Waltz. Energies above 90 (dashed line) are considered amyloidogenic. A: 5898_B01 (FG loop sequence "KMRDYR"), B: 5898_C01 FG loop sequence "SLRDYG") , C: 5898_F01 (FG loop sequence "MSRDYG"), D: 6199_B01 (FG loop sequence "GSRDYE"), E: 6199_D08 (FG loop sequence "VLRDYR"), F: 6199_A07 (FG loop sequence "LLRDYV"). Figure produced using Mathematica™.

Waltz output is a measure of confidence that a given segment is amyloidogenic. A cut-off value must be selected to establish the desired sensitivity/selectivity trade-off (i.e. to maximize detection of amyloidogenic segments and minimize the number of false positives). The all-or-nothing nature of this measure of amyloidogenicity (Figure 4.9) and the high identity of the Adnectin™ sequences combine to give nearly identical scores for clusters of Adnectins™ (Figure 4.10). There is a positive correlation between the average Waltz scores and the Adnectin™ IB formation data, but like the 3D Profile and TANGO results, the scatter is substantial.

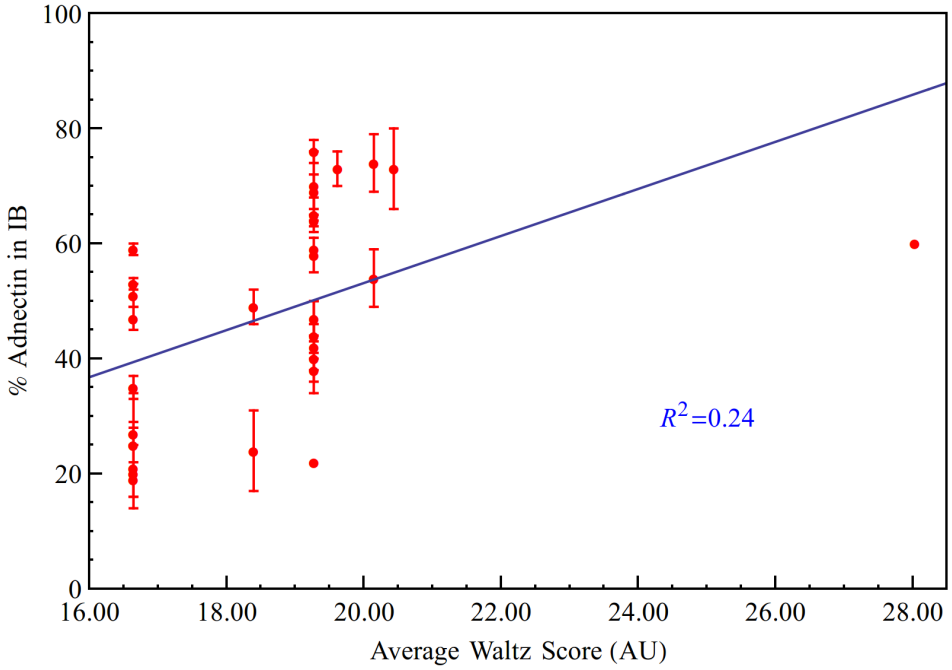

Figure 4.10: % Adnectin™ in IBs at 4 hours post-induction (average of two growths; error bars indicate range) vs. average segment scores generated using Waltz. Blue line: linear fit of the data. Figure produced using Mathematica™.

PASTA 2.0

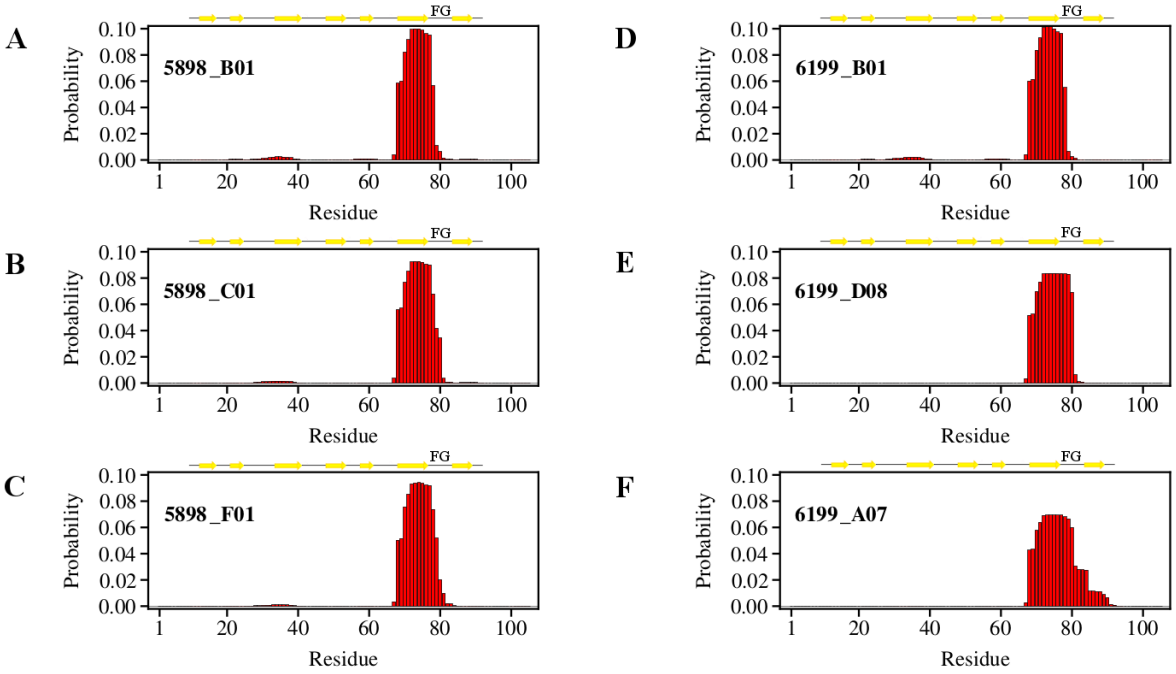

Figure 4.11: Adnectin™ sequence segment aggregation propensity profile produced using PASTA 2.0. A: 5898_B01 (FG loop sequence "KMRDYR"), B: 5898_C01 FG loop sequence "SLRDYG") , C: 5898_F01 (FG loop sequence "MSRDYG"), D: 6199_B01 (FG loop sequence "GSRDYE"), E: 6199_D08 (FG loop sequence "VLRDYR"), F: 6199_A07 (FG loop sequence "LLRDYV"). Figure produced using Mathematica™.

The PASTA 2.0 algorithm outputs the normalized probability of fibril formation for each residue. The aggregation propensity profiles in Figure 4.11 highlight the segment including β-strand F as most likely to form fibrils, but the main differences between Adnectin™ profiles lie in the segments containing FG loop residues. The average of these normalized probabilities is the same for each Adnectin™; instead of IB data vs. average score, Figure 4.12 shows a plot of IB formation data vs. the PASTA 2.0 energy score of the most aggregation-prone sequence segment pairing. The R^2 of the linear fit is better than the sliding window methods considered above, but many Adnectins™ have segment pairings with similar PASTA 2.0 energy scores, despite a range of different IB formation results.

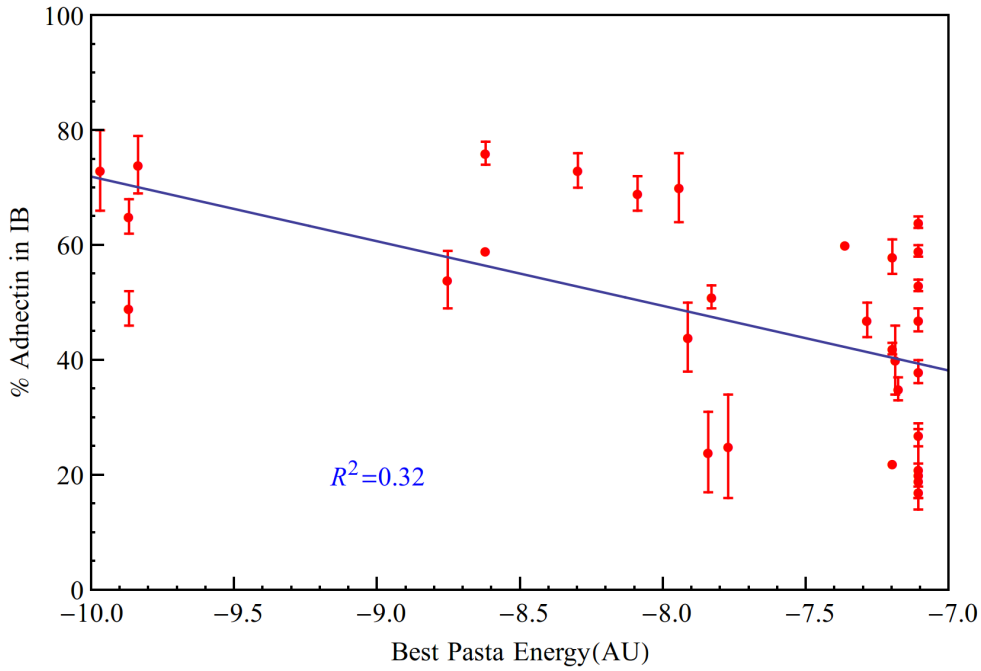

Figure 4.12: % Adnectin™ in IBs at 4 hours post-induction (average of two growths; error bars indicate range) vs. lowest (most aggregation-prone) segment score generated using PASTA 2.0. Blue line: linear fit of the data. Figure produced using Mathematica™.

AGGRESCAN

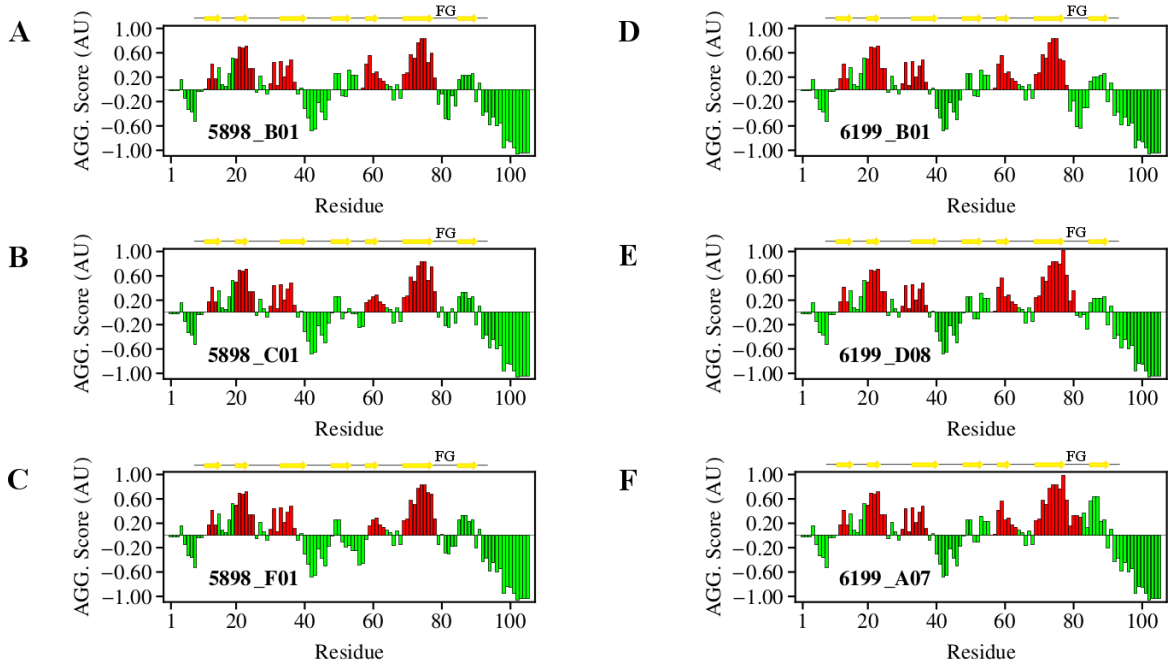

Figure 4.13: Adnectin™ sequence segment aggregation propensity profile produced using AGGRESCAN. Five consecutive residues with scores greater than -0.02 constitute a "hot spot" (red). A: 5898_B01 (FG loop sequence "KMRDYR"), B: 5898_C01 FG loop sequence "SLRDYG") , C: 5898_F01 (FG loop sequence "MSRDYG"), D: 6199_B01 (FG loop sequence "GSRDYE"), E: 6199_D08 (FG loop sequence "VLRDYR"), F: 6199_A07 (FG loop sequence "LLRDYV"). Figure produced using Mathematica™.

The AGGRESCAN aggregation propensity profiles identify several regions of interest (Figure 4.13), and the algorithm's predictions must be taken seriously; the correlation of Adnectin™ IB formation with sequence-average AGGRESCAN scores is clearly significant (Figure 4.14). The aggregation propensity of β-strand F is high in all of the profiles, though modulated by the composition of the FG loop. The high scores assigned to β-strand B are also noteworthy, but are equal for all of the Adnectins™; the β-strand B scores cannot have contributed to AGGRESCAN's ability to discriminate between Adnectins™ with different IB formation propensities.

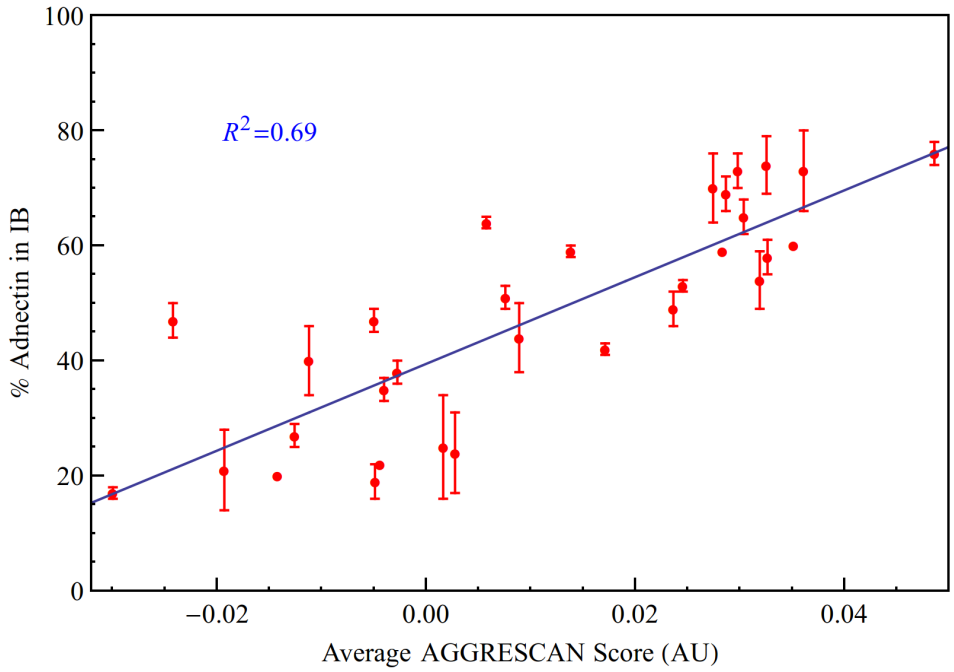

Figure 4.14: % Adnectin™ in IBs at 4 hours post-induction (average of two growths; error bars indicate range) vs. average segment scores generated using AGGRESCAN. Blue line: linear fit of the data. Figure produced using Mathematica™.

Zyggregator

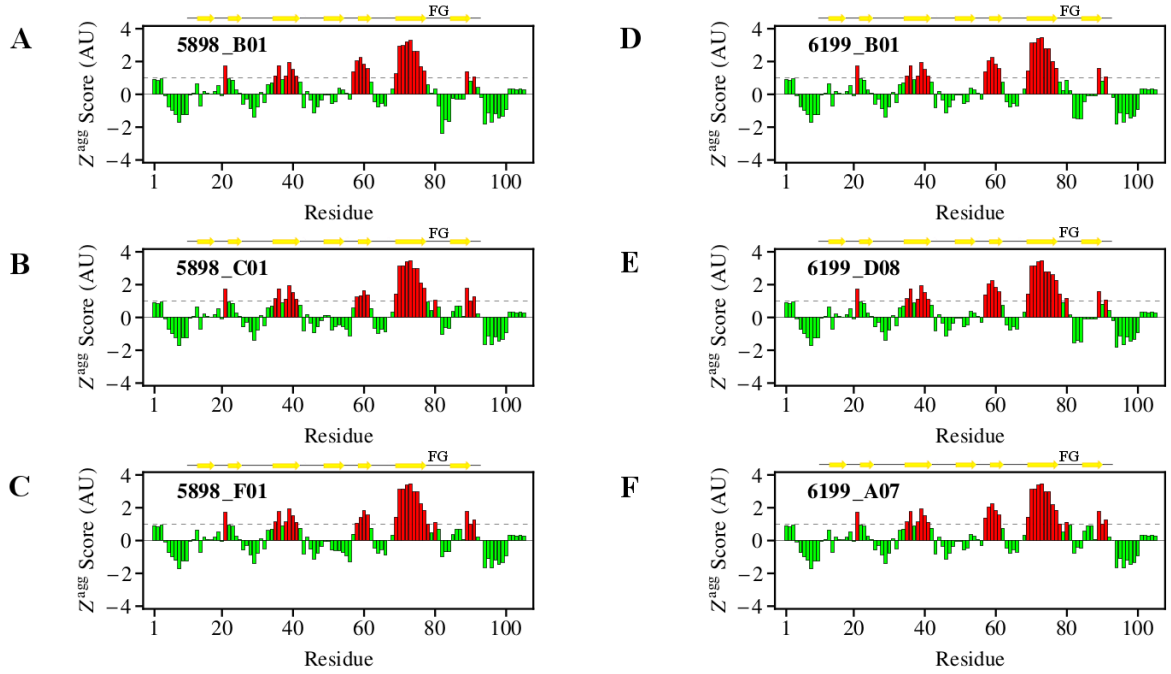

Figure 4.15: Adnectin™ sequence segment aggregation propensity profile produced using Zyggregator. Scores greater than 1 (red) are considered aggregation-prone. A: 5898_B01 (FG loop sequence "KMRDYR"), B: 5898_C01 FG loop sequence "SLRDYG") , C: 5898_F01 (FG loop sequence "MSRDYG"), D: 6199_B01 (FG loop sequence "GSRDYE"), E: 6199_D08 (FG loop sequence "VLRDYR"), F: 6199_A07 (FG loop sequence "LLRDYV"). Figure produced using Mathematica™.

The Z^{agg} score is normalized such that it is greater than unity if the aggregation propensity at a given position is one standard deviation more aggregation-prone than a random sequence. Scores greater than unity are concentrated in the vicinity of β-strand F (Figure 4.15).

The correlation of IB formation data with average Z^{agg} scores (Figure 4.16) is ostensibly lower than that achieved by AGGRESCAN (Figure 4.14), but when the variation in the IB data (error bars; Figures 4.14 and 4.16) is considered, the performances of the two algorithms seem very similar.

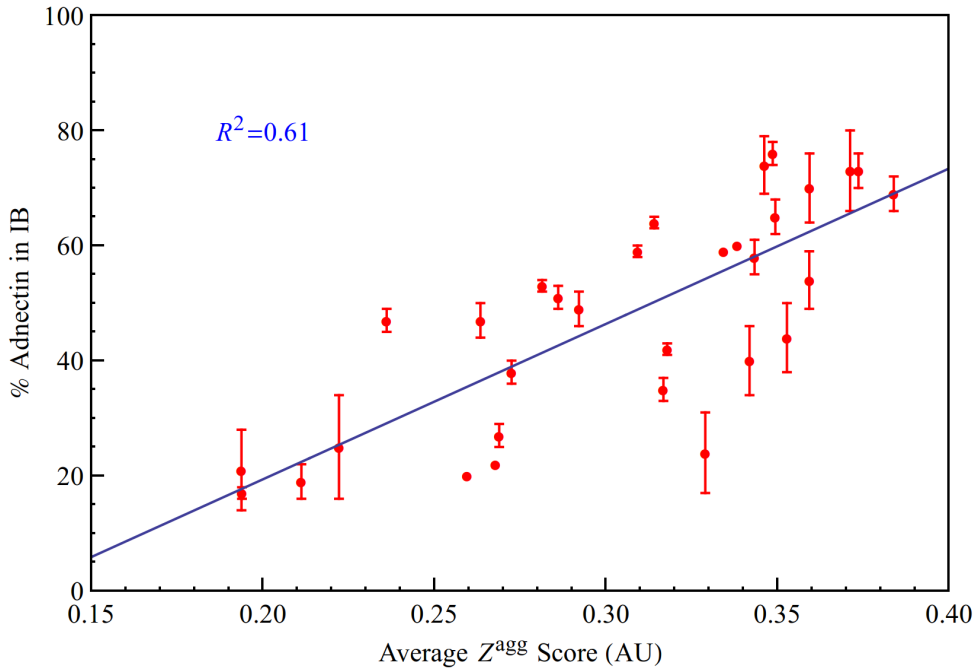

Figure 4.16: % Adnectin™ in IBs at 4 hours post-induction (average of two growths; error bars indicate range) vs. Z^{agg} score generated using Zyggregator. Blue line: linear fit of the data. Figure produced using Mathematica™.

FoldAmyloid

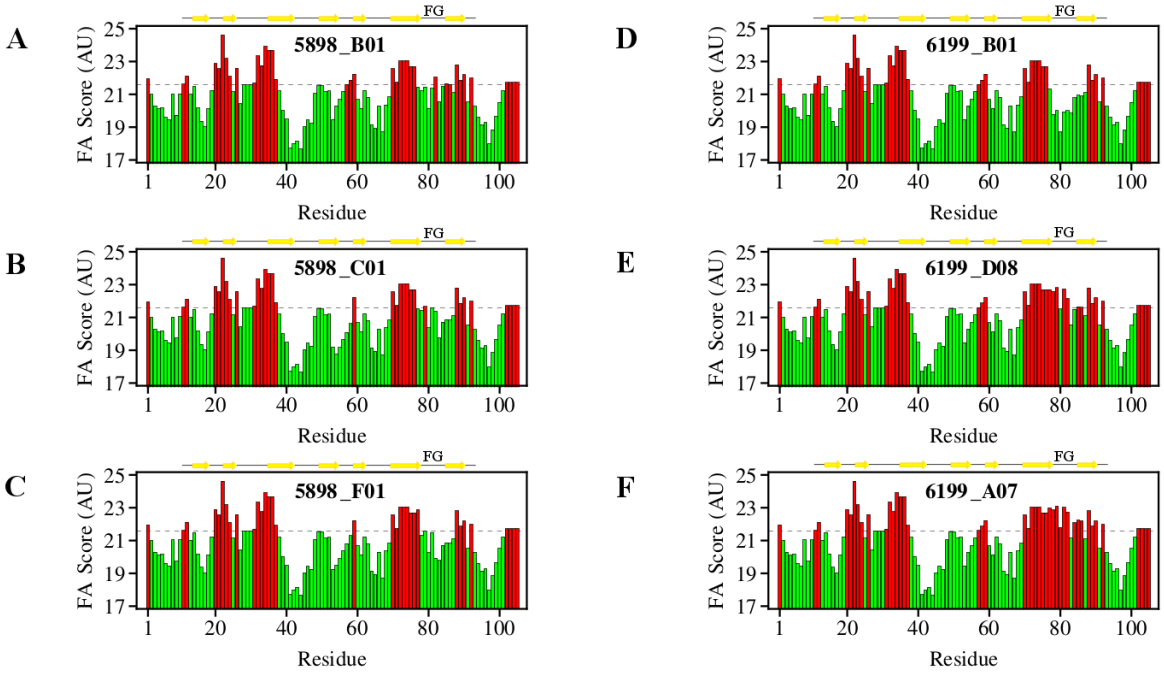

Figure 4.17: Adnectin™ sequence segment aggregation propensity profile produced using FoldAmyloid. Scores greater than 21.4 (red) are considered amyloidogenic. A: 5898_B01 (FG loop sequence "KMRDYR"), B: 5898_C01 FG loop sequence "SLRDYG") , C: 5898_F01 (FG loop sequence "MSRDYG"), D: 6199_B01 (FG loop sequence "GSRDYE"), E: 6199_D08 (FG loop sequence "VLRDYR"), F: 6199_A07 (FG loop sequence "LLRDYV"). Figure produced using Mathematica™.

The FoldAmyloid predictions presented here are based upon the expected packing density of residues. Of the methods that we have applied to the Adnectin™ sequences, this is the only one that explicitly includes packing density; however, packing density may be implicitly incorporated into the measures of β-sheet and/or β-aggregate propensity used in other methods (note that many of the FoldAmyloid scores that exceed the amyloidogenic threshold of 21.4 in Figure 4.17 are located in a β-strand).

FoldAmyloid is an elegant method, based exclusively upon expert knowledge culled from high-resolution structures in the PDB, and its ability to discriminate between the Adnectins™ (Figure 4.18) is comparable to those of AGGRESCAN (Figure 4.14) and Zyg-

gregator (Figure 4.16).

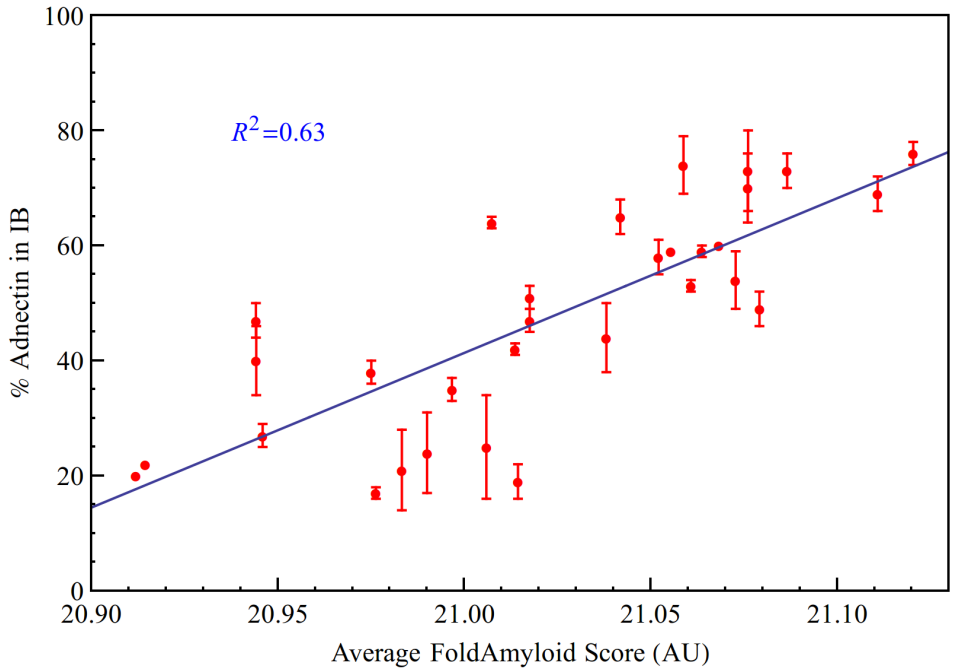

Figure 4.18: % Adnectin™ in IBs at 4 hours post-induction (average of two growths; error bars indicate range) vs. average score generated using FoldAmyloid. Blue line: linear fit of the data. Figure produced using Mathematica™.

4.4.3 Discussion

The significant correlations found between the AGGRESCAN, Zyggregator, and FoldAmyloid predictions and the Adnectin™ IB formation data is a strong validation of their use for this application. The lower correlations achieved by the other sliding window methods indicate that they do not discriminate well between Adnectins™ with varying IB formation propensity, but they may still succeed in identifying aggregation-prone regions that the Adnectins™ have in common. Accordingly, all of the sliding window methods surveyed in this section were included in a consensus aggregation propensity profile of Adnectin™ 6199_A07 (selected on the basis of its high IB formation propensity) (Figure 4.19).

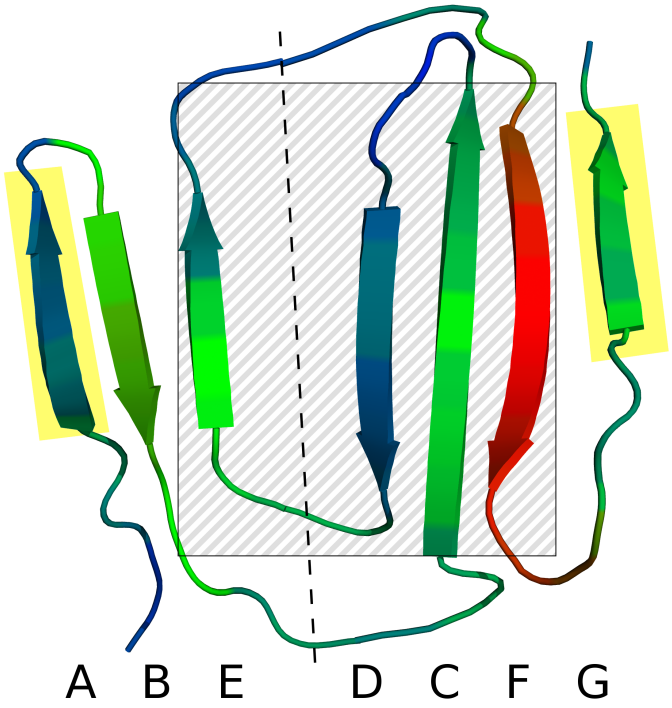

Figure 4.19: Flat Adnectin™ representation coloured by consensus aggregation propensity from blue (low; 0.4) to red (high; 6.5). β-strands identified by letter, beginning at the N terminus. Dashed line divides the two halves of the β-sandwich. Striped box indicates the strands that form the folding nucleus. First strands to lose structure upon unfolding due to mechanical stress highlighted in yellow. Figure produced using PyMOL and GIMP.

The sequence position-specific aggregation propensity scores from each method were normalized such that they span a range from 0 (lowest predicted aggregation propensity) to 1 (highest predicted aggregation propensity) and summed on an equal-weight basis. The flat Adnectin™ representation in Figure 4.19 was coloured from blue (low consensus aggregation propensity) to red (high consensus aggregation propensity). The β-strands indicated by the striped box (Figure 4.19) form the "folding nucleus" (they contain residues thought to be structured in the transition state) [113]. It is possible that the denatured state ensemble is energetically biased toward average structure with native-like characteristics resembling those of the folding nucleus; this could reduce the exposure of the segments that comprise the folding nucleus, giving them fewer opportunities to initiate aggregation (see Chapter 1). The β-strands highlighted in yellow (Figure 4.19) are those thought to be the first to lose structure during mechanical unfolding (Section 1.2). If high-energy conformations are transiently sampled in the native state ensemble, it is possible that these strands (A, G) are more exposed than the other strands, and therefore more available to initiate aggregation from a native-like state. Additional support for the idea that the A and G β-strands are more mobile than the rest can be found in the discussion of Fn3 structural dynamics (H-D exchange; B-factors) in Section 1.2.

The consensus aggregation propensity profile (Figure 4.19) makes clear that the F β-strand has high potential to aggregate by many different measures. The A and G β-strands are potentially more available to aggregate, but their propensity to do so is predicted to be lower. This is consistent with the hypothesis that edge strands (which are solvent exposed even in the native state) may be under selective pressure for low aggregation propensity [154]. The aggregation-prone F β-strand becomes an edge strand if the G strand is out of position. The B strand is predicted by only some of the algorithms to have high potential to aggregate (hence its moderate consensus aggregation propensity), but a peptide with a sequence that corresponds to that of the B strand in wild-type ^{10}Fn3 (SLLISWD) has been shown to help initiate fibronectin fibrillogenesis [101].

Chapter 5

Modelling of Adnectin™ Structures

5.1 Introduction

It is widely believed that, in principle, the native structure of a protein can be determined solely on the basis of its amino acid sequence [155]. In practice, accurate *de novo* structure prediction from amino acid sequence remains elusive for all but the smallest, single-domain proteins. Fortunately, as a result of exponential growth in the number of high-resolution structures available in the PDB [156], this limitation can often be circumvented. Estimates of the number of unique protein folds range from ∼450 to ∼10000 [157]; though these numbers vary by orders of magnitude, the differences between them seem insignificant when compared to the more than 13 million protein sequences in the National Center for Biotechnology Information's curated, non-redundant database [158]. As of 2009, approximately one quarter of the single-domain sequences in this database could be associated with a structure (by membership in a family of similar sequences, the structure of at least one of which is known) [159].

The process of constructing a structural representation of a target protein using a homologous protein of known structure as a template is called homology (or comparative) modelling. For long sequence alignments, structural homology can be inferred with confidence if the target protein sequence is greater than 40% identical to the template sequence [160]. Because the Adnectins™ differ from wild-type ^{10}Fn3 (for which several high-resolution structures are available in the PDB [110, 161, 162]) only in three loops, this threshold is exceeded by a wide margin.

Nevertheless, modelling Adnectin™ structures remains a challenge; dealing with flexible loops is one of the most difficult parts of homology modelling [163, 164], and while the

high sequence identity of the targets with potential templates solves one problem, minimal variation between Adnectins™ creates another. In order to be useful, the models must capture subtle structural differences.

The accuracy of a homology model can only be verified by comparison with experimental data. Modelling methods can be evaluated by choosing targets for which high-resolution structures are available for comparison, but the true value of homology modelling lies in the creation of high-resolution models of proteins for which no experimentally determined structures exist.

Using a template based upon the ^{10}Fn3 domain from PDB structure 1FNF [110], we have employed the kinematic loop modelling features of the Rosetta software package [165] to generate an ensemble of models for each Adnectin™ (Appendix A). We will attempt to validate these models by comparing scores assigned by the Rosetta energy function to Adnectin™ IB formation data (Chapter 2) and DSC data (Chapter 3).

5.1.1 Protein Flexibility and Ensemble Properties

Proteins are dynamic molecules under physiologically relevant conditions, and in terms of the energy landscape theory of protein folding are best characterized as occupying a native basin rather than a single state (Chapter 1). The conformational flexibility of proteins is well-known, and is partly captured in both crystal and solution NMR structural data as a degree of uncertainty about the relative positions of atoms. Both methods of structure determination average the properties of numerous molecules; the positions of atoms in flexible loops and near the termini may be poorly defined if these regions adopt different conformations in otherwise identical molecules. In crystal structures, the uncertainty of each atomic position (which may be influenced by experimental limitations, in addition to the existence of different conformations) is quantified by the associated B-factor (Section 1.2). Alternatively, the conformational heterogeneity of the native basin of an energy landscape can be described using an ensemble of structures [166].

5.1.2 Kinematic Closure

The kinematic closure (KIC) method of loop modelling, as implemented in Rosetta [165], is capable of generating an ensemble of realistic loop conformations because it uses knowledge derived from high-resolution protein structures deposited in the PDB to determine as many of the loop residue dihedral angles as possible. During each iteration of the KIC algorithm, the C_α carbons of three nonconsecutive residues are selected as "pivots". The torsion angles

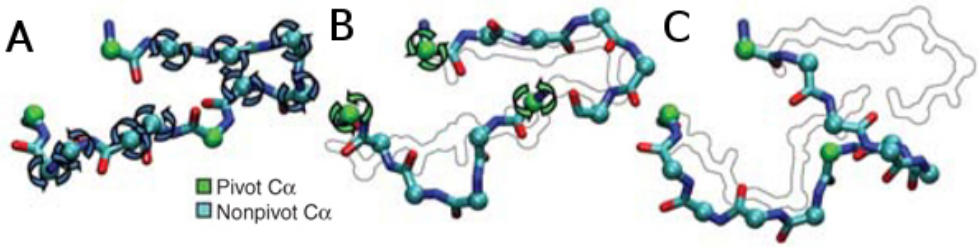

Figure 5.1: Illustration of the kinematic closure loop modelling process; A: pivot residues are selected (green), B: non-pivot torsion angles are replaced, opening the loop, C: pivot residue torsion angles that close the loop are determined. Figure reproduced from [165].

of all non-pivot residues are then sampled according to residue-specific Ramachandran probabilities [167], and N-C_α-C bond angles are set to random values less than one-half of a standard deviation from the mean of angles observed in sub-1.0Å resolution structures from the PDB. As these initial steps have the effect of opening the loop, pivot residue torsion and bond angles that close it are then determined by solving a system of equations derived from the constraints imposed by the loop closure problem.

5.1.3 Adnectin™ FG Loop Modelling

The Adnectin™ sequences (Appendix A) differ from each other primarily in the FG loop, but differ from the wild-type ^{10}Fn3 sequence in the BC, DE, and FG loops. We elected to focus on modelling just the Adnectin™ FG loop sequences; an exponentially larger number of models would have been required to describe all possible combinations of BC/DE/FG loop conformations, and since the BC and DE loops do not directly interact with the FG loop (Figure 5.2), it is unlikely that modelling their conformational variability would provide additional basis for distinguishing between Adnectins™.

5.2 Methods

The Adnectin™ homology models were constructed using PyMOL, FoldX, and Rosetta. The wild-type ^{10}Fn3 domain (residues 1418-1509) was extracted from the 1FNF PDB structure, and the FG loop shortened from ten to six residues using PyMOL (the resulting loop conformation is invalid, but will subsequently be replaced by the loop modelling

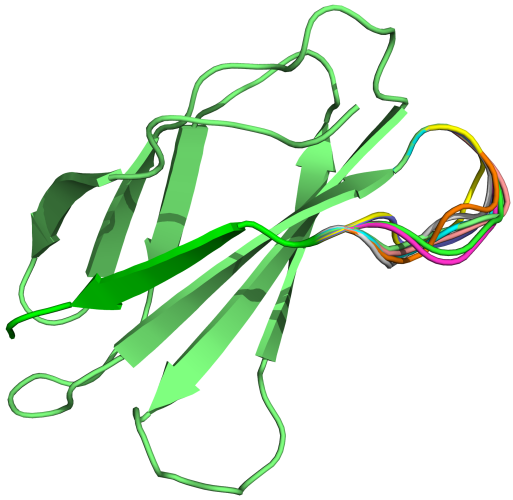

Figure 5.2: Model of Adnectin™ 6199_D06 showing nine different FG loops (constructed using the KIC method) superimposed. Figure produced using PyMOL.

procedure). FoldX was then used to create a template for each Adnectin™ (in PDB format) by mutating the six remaining residues of the FG loop to the residues highlighted in Appendix A), as described in [168].

The Rosetta KIC loop modelling protocol [165] was executed on the "orca" cluster of the Shared Hierarchical Academic Research Computer Network (SHARCNET), using the following options:

-loops:remodel perturb_kic

-loops:refine refine_kic

-in:file:fullatom

-ex1

-ex2

Coefficients for the linear combination of Zyggregator (Z^{agg}) scores and ensemble average Rosetta scores (plotted against IB formation at 4 h post-induction) in Figure 5.5 were determined using a quasi-Newton method (as implemented in the Microsoft Excel 2007 Solver add-in) to maximize the R^2.

5.3 Results & Discussion

The Rosetta energy function reports scores in arbitrary units that we will refer to as Rosetta Energy Units (REU). For each Adnectin™ FG loop sequence, ten groups of one thousand models were created using the Rosetta KIC protocol (for a total of ten thousand models per Adnectin™), and scored by the Rosetta energy function. The standard deviations listed in Table 5.1 are based on the ten average scores (one for each group of one thousand models) per mutant, and are intended to show that one thousand models is generally enough to determine a representative average score. Adnectin™ 6199_D05 is the only exception (i.e. the standard deviation of ensemble average scores is high, suggesting that more than one thousand models may be required in order to determine a representative average score for this mutant).

The Rosetta energy function is intended to provide a measure of thermodynamic stability [151]; the lack of a strong correlation between the DSC data (Chapter 3) and the scores in Table 5.1 is initially counterintuitive (Figure 5.3). This finding may be rationalized by considering that the differences between the Adnectins™ are largely restricted to a single flexible loop, and that this loop is within a part of the native fold that may be unusually dynamic (Chapter 1).

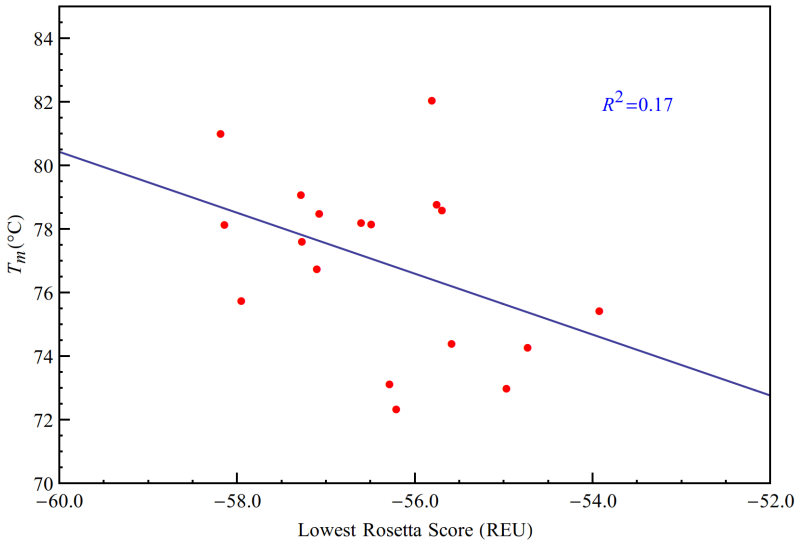

Figure 5.3: T_m's determined by DSC vs. lowest Rosetta score from each Adnectin™ ensemble of structures. Figure produced using Mathematica™.

Table 5.1: **Loop model Rosetta energy scores; standard deviations based upon ten average scores (one for each group of 1000 models).**

Adnectin™	**FG Loop Sequence**	**Lowest Score (REU)**	**Average Score (REU)**	**Standard Deviation**
5898_B01	KMRDYR	-54.97	-51.05	0.10
5898_C01	SLRDYG	-58.01	-53.97	0.12
5898_C02	LLRDYG	-57.10	-51.56	0.59
5898_E01	SLRDYA	-55.75	-50.79	0.13
5898_F01	MSRDYG	-57.27	-52.56	0.13
5898_H01	NLRDYG	-57.08	-53.48	0.18
6199_A03	KVRDYR	-56.21	-50.47	0.28
6199_A05	YLRDYT	-54.61	-48.78	0.15
6199_A07	LLRDYV	-54.54	-49.67	0.10
6199_B01	GSRDYE	-58.19	-55.48	0.12
6199_B02	TQRDYG	-58.14	-53.07	0.17
6199_B03	TWRDYL	-56.60	-51.03	0.14
6199_B04	CRRDYG	-56.43	-53.19	0.07
6199_B05	EMRDYG	-56.45	-52.33	0.12
6199_B07	ERRDYR	-56.49	-51.45	0.12
6199_C05	LVRDYG	-55.96	-49.85	0.30
6199_D01	RIRDYG	-56.29	-51.60	0.10
6199_D05	FIRDYG	-55.45	-49.01	1.78
6199_D06	SRRDYG	-57.96	-53.90	0.13
6199_D07	ALRDYV	-57.28	-53.02	0.08
6199_D08	VLRDYR	-53.92	-49.00	0.06
6199_E01	RSRDYR	-57.80	-52.34	0.11
6199_E02	HFRDYG	-56.19	-53.18	0.13
6199_E03	KLRDYL	-55.59	-51.34	0.12
6199_E06	RLRDYE	-54.73	-50.99	0.10
6199_F01	DYRDYL	-55.81	-51.77	0.17
6199_F07	SLRDYV	-56.60	-52.26	0.12
6199_F08	TLRDYM	-55.70	-50.74	0.10
6199_G07	LIRDYG	-56.43	-49.90	0.27
6199_H04	LFRDYG	-56.69	-52.09	0.10
6199_H07	QLRDYS	-55.07	-49.94	0.13

We have shown that mutations to residues in this flexible loop can change the T_m by almost 10°C (Table 3.1), but because the loop is solvent-exposed in the native structure, the mutations may increase or decrease the energies of the native and denatured state ensembles in tandem (with little change in the ΔG between them; Figure 1.2). The Rosetta energy function partially compensates for this phenomenon; each amino acid type is assigned a "reference" energy value, and the energy of the denatured state is approximated by summing these values over the whole sequence [150]. Such compensation implicitly assumes that there are no interactions between residues in denatured conformations; we suspect that breakdowns in this assumption contribute to the poor correlation between the T_m values and the Rosetta scores.

The comparison between the ensemble loop model scores and the Adnectin™ IB formation data proved more promising (Figure 5.4). Interestingly, equal-weight averages of loop model ensemble scores correlate much better with the IB formation data than either the lowest scores of each ensemble, or averages weighted according to probabilities determined using Equation 1.2. This could be evidence that the conformations most pertinent to IB formation are not those with the lowest energy.

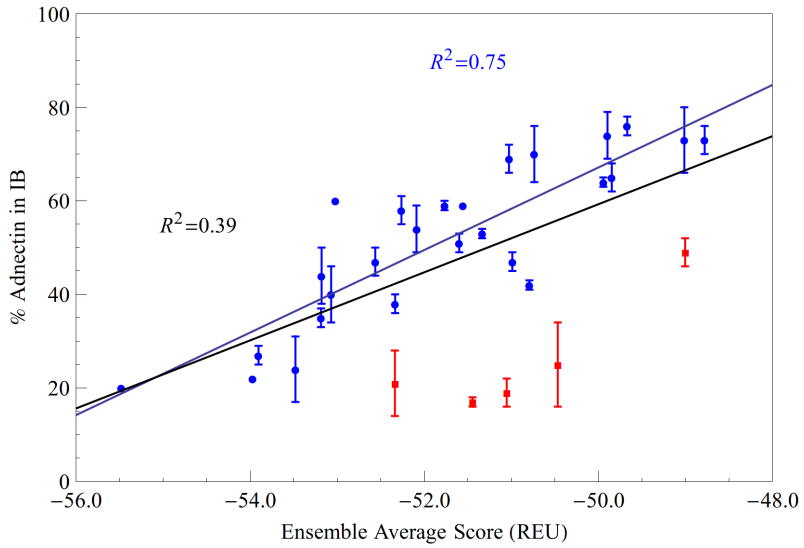

Figure 5.4: Percentage of Adnectins™ found in IBs at 4 hours post-induction (average of two growths; error bars indicate range) vs. ensemble average Rosetta scores. Red squares: Adnectins™ with arginine in the sixth position of the FG loop. Blue circles: All other Adnectins™. Black line: linear fit of all data points. Blue line: linear fit of blue points only. Figure produced using Mathematica™.

The correlation of the ensemble average loop model scores with IB formation is high enough to be intriguing (Figure 5.4; black line), but low enough to make it clear that the predictive value of the loop models has limits; other factors must contribute to Adnectin™ IB formation. The linear relationship was found to be much stronger if a group of outliers, all of which have an arginine in the sixth position of the FG loop (Figure 5.4; red squares), were excluded from the fit. This suggests that having an arginine in this position discourages IB formation in a way that is not captured by the loop models.

The implications of the correlation between the ensemble average Rosetta scores and IB formation are not entirely clear; however, the fact that models of native structure can be used to predict aggregation (IB formation) suggests the hypothesis that the models capture favourable interactions that restrict the exposure of aggregation-prone sequence segments. If the scores can be interpreted as a measure of exposure, better predictions of IB formation should be possible through combination with a measure of the intrinsic aggregation propensity of the (denatured) sequences (Chapter 4).

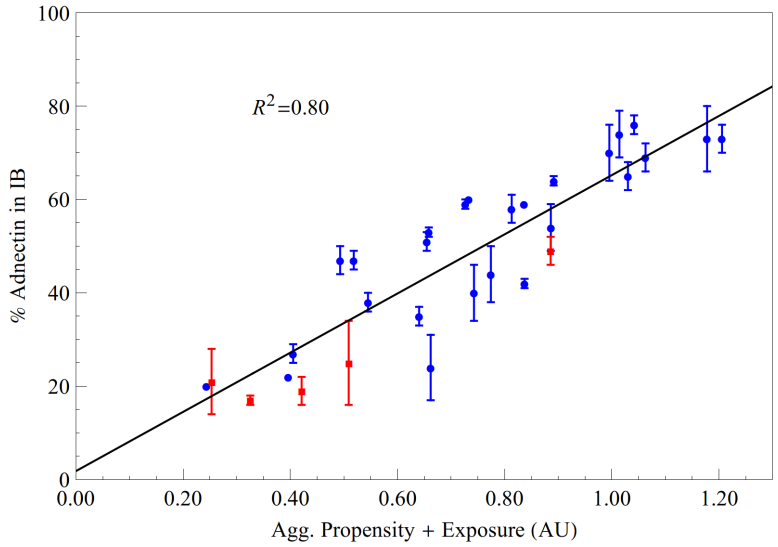

Figure 5.5: Percentage of Adnectins™ found in IBs at 4 hours post-induction (average of two growths; error bars indicate range) vs. a linear combination of normalized Z^{agg} and ensemble average Rosetta scores. Red squares: Adnectins™ with arginine in the sixth position of the FG loop. Blue circles: All other Adnectins™. Black line: linear fit of all data points. Figure produced using Mathematica™.

The Chiti-Dobson equation (Section 4.3), AGGRESCAN (Section 4.4.2), Zyggregator (Section 4.4.2), and FoldAmyloid (Section 4.4.2) all generated aggregation predictions that

correlate significantly with the Adnectin™ IB formation data (Chapter 2). Of these, we chose Zyggregator as the measure of aggregation propensity to be included in the combination, because the Z^{agg} scores have the lowest correlation with the ensemble average loop model scores.

The measure of aggregation propensity and the measure of exposure to be combined would ideally be independent variables; in this instance, because the same mutations may simultaneously increase or decrease both aggregation propensity and exposure, the square of the correlation coefficient between the Z^{agg} scores and the ensemble average Rosetta scores (R^2_{p-e}) is non-zero (equal to 0.065). The R^2 achievable through a linear combination of these non-independent variables is given by Eq. 5.1 [169], where values for the square of the correlation coefficient between the Z^{agg} scores and the Adnectin™ IB formation data (R^2_p; 0.61) and the square of the correlation coefficient between the ensemble average Rosetta scores and the Adnectin™ IB formation data (R^2_e; 0.39) have been taken from Figures 4.16 and 5.4, respectively. Allowing for rounding error, this is indeed the R^2 of the linear regression shown in Figure 5.5. The Adnectins™ with arginine in the sixth position of the FG loop (red outliers in Figure 5.4) fall close to the line of best fit in Figure 5.5.

$$R^2_{combination} = \frac{R^2_p + R^2_e - 2 \cdot R_p \cdot R_e \cdot R_{p-e}}{1 - R^2_{p-e}} = 0.80 \tag{5.1}$$

Chapter 6

Conclusions & Future Work

6.1 Conclusions

Inclusion body formation is a complex phenomenon, influenced by numerous factors (Chapter 1). In order to study the molecular mechanisms governing IB formation in *E. coli*, different sequences may be expressed; however, as the number of differences grows, it becomes increasingly difficult to determine how the many factors influencing IB formation could be impacted. The Adnectins™ (Appendix A) comprise an excellent model system for the study of IB formation because the proportion of total Adnectin™ found in IBs varies considerably despite amino acid differences at just three positions within a single flexible loop (Chapter 2). The common core shared by all Adnectins™ is identical to that of wild-type ^{10}Fn3, for which several high-resolution structures are available (Chapter 1), providing a solid foundation for the modelling of Adnectins™ (Chapter 5).

Some of the variation in Adnectin™ IB formation can be attributed to differences in the intrinsic aggregation propensity of their sequences. This is evident from the moderately high correlation of IB formation with sequence-based predictions generated using the Chiti-Dobson equation, AGGRESCAN, Zyggregator, and FoldAmyloid (Chapter 4), and also from the variable pH-dependent association/aggregation behaviour of thermally denatured Adnectins™ (Chapter 3). The accuracy of IB formation predictions can be improved by considering not only the intrinsic aggregation propensity of sequence segments, but also the degree to which they may be exposed. Adnectin™ IB formation correlates poorly with experimentally determined global stability (Chapter 3), but very well with a subset of the average energy scores of Adnectin™ homology model ensembles (Chapter 5). The ensemble average scores may be capturing subtle shifts in the energetic bias toward native structure

that restricts the exposure of aggregation-prone sequence segments. Consistent with this hypothesis, the most accurate Adnectin™ IB formation predictions were obtained using a linear combination of aggregation propensity (calculated using Zyggregator) and ensemble average energy scores (as a measure of aggregation-prone sequence segment exposure).

6.2 Future Work

6.2.1 Logical Extensions

Study Additional Adnectins™

Many (∼200) different Adnectin sequences are available in the library provided by Adnexus (a division of Bristol-Meyers Squibb); so far, we have characterized a subset carefully selected for similarity of sequence, and established a framework for understanding IB formation differences within this subset in terms of the exposure and intrinsic aggregation propensity of sequence segments. Additional Adnectins™ with greater diversity in their FG loop sequences could be studied in order to determine how well this framework generalizes.

Dynamic Light Scattering

The exotherms observed by DSC upon thermal denaturation of Adnectins™ (pH and protein concentration dependent) have been interpreted as evidence of association and/or aggregation (Chapter 3). It would be useful to confirm the validity of this interpretation, and to characterize the size distribution of the resulting particles, for example, using dynamic light scattering (DLS) to monitor the thermal denaturation of Adnectins™.

Adnectin™ Folding Kinetics

In the absence of a strong relationship between Adnectin™ IB formation and global stability, the correlation observed between IB formation data and the energy scores assigned to homology models of natively folded Adnectin™ (Chapter 5) could be interpreted as evidence that local unfolding of native structure leads to IB formation through the exposure of aggregation-prone sequence segments. However, native-like structure may also be found in the DSE or intermediate ensembles (Section 1.1.3). If aggregation from these ensembles contributes to IB formation, folding rates and/or evidence for intermediates may vary between Adnectins™.

Loop Model Ensemble Analysis

The fact that Adnectin™ IB formation data correlates better with equal-weight averages of ensemble energy scores than it does with the lowest scores or with averages weighted by probabilities determined using Eq. 1.2 suggests that the conformations most relevant to IB formation are not those with the lowest energy. Identifying these relevant conformations presents a challenge. Clustering the conformations in each ensemble by RMSD, and looking for clustering patterns common to different Adnectins™ with similar IB formation propensities may yield some clues.

6.2.2 New Directions: IB Structure and Stability

The Adnectins™ differ primarily in the FG loop, and the methods employed to explore differences in IB formation have been focused on this region. Despite the success of this narrow focus, other parts of the proteins could be involved in IB formation, and there remains much to learn about the structure and stability of Adnectin™ IBs.

Absorption spectra obtained using Fourier transform infrared spectroscopy (FTIR), particularly in the Amide I region (1600-1700 cm^{-1}), could be used to study and compare the secondary structure present in natively folded Adnectins™ and IBs [37]. We have studied many natively folded Adnectins™ using DSC; the same technique could be applied to Adnectin™ IBs in order to determine whether or not there are measurable differences in IB thermostability. Also, protease digestion of Adnectin™ IBs, coupled with analysis of the fragments by mass spectrometry, could be used to discern which sequence segments of the constituent proteins are involved in stable intermolecular contacts, and which are less protected.

APPENDICES

Appendix A

Adnectin™ Amino Acid Sequences

Table A.1: **Adnectin™ amino acid sequences. FG loop highlighted.**

Identifier	Sequence
5898_B01	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT KMRDYR PISINYRTEIDKPSQHHHHHH
5898_C01	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KGKYTATISGLKPGVDYTITVYAVT SLRDYG PISINYRTEIDKPSQHHHHHH
5898_C02	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT LLRDYG PISINYRTEIDKPSQHHHHHH
5898_E01	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT SLRDYA PISINYRTEIDKPSQHHHHHH
5898_F01	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KDRYTATISGLKPGVDYTITVYAVT MSRDYG PISINYRTEIDKPSQHHHHHH
5898_H01	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT NLRDYG PISINYRTEIDKPSQHHHHHH
6199_A03	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT KVRDYR PISINYRTEIDKPSQHHHHHH
6199_A05	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT YLRDYT PISINYRTEIDKPSQHHHHHH
6199_A07	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT LLRDYV PISINYRTEIDKPSQHHHHHH

Identifier	Sequence
6199_B01	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT GSRDYE PISINYRTEIDKPSQHHHHHH
6199_B02	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT TQRDYG PISINYRTEIDKPSQHHHHHH
6199_B03	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT TWRDYL PISINYRTEIDKPSQHHHHHH
6199_B04	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT CRRDYG PISINYRTEIDKPSQHHHHHH
6199_B05	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT EMRDYG PISINYRTEIDKPSQHHHHHH
6199_B07	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT ERRDYR PISINYRTEIDKPSQHHHHHH
6199_C05	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT LVRDYG PISINYRTEIDKPSQHHHHHH
6199_D01	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT RIRDYG PISINYRTEIDKPSQHHHHHH
6199_D05	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT FIRDYG PISINYRTEIDKPSQHHHHHH
6199_D06	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT SRRDYG PISINYRTEIDKPSQHHHHHH
6199_D07	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT ALRDYV PISINYRTEIDKPSQHHHHHH
6199_D08	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT VLRDYR PISINYRTEIDKPSQHHHHHH
6199_E01	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT RSRDYR PISINYRTEIDKPSQHHHHHH
6199_E02	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT HFRDYG PISINYRTEIDKPSQHHHHHH
6199_E03	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT KLRDYL PISINYRTEIDKPSQHHHHHH
6199_E06	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT RLRDYE PISINYRTEIDKPSQHHHHHH

Identifier	Sequence
6199_F01	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT DYRDYL PISINYRTEIDKPSQHHHHHH
6199_F07	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT SLRDYV PISINYRTEIDKPSQHHHHHH
6199_F08	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT TLRDYM PISINYRTEIDKPSQHHHHHH
6199_G07	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT LIRDYG PISINYRTEIDKPSQHHHHHH
6199_H04	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT LFRDYG PISINYRTEIDKPSQHHHHHH
6199_H07	MGVSDVPRDLEVVAATPTSLLISWSARLKVARYYRITYGETGGNSPVQEFTVP KNVYTATISGLKPGVDYTITVYAVT QLRDYS PISINYRTEIDKPSQHHHHHH

Appendix B

Differential Scanning Calorimetry Data

Table B.1: **Differential scanning calorimetry data at pH 4.0; ordered and highlighted to match Table 3.1.**

	Lower Concentration			Higher Concentration		
Adnectin™	**Conc. (mg/mL)**	**T_m (°C)**	**ΔH_{VH} (kcal/mol)**	**Conc. (mg/mL)**	**T_m (°C)**	**ΔH_{VH} (kcal/mol)**
6199_B07	0.22	78.0	82.8	0.46	78.3	81.1
5898_B01	0.17	73.2	75.4	0.41	72.8	76.8
6199_B01	0.27	81.0	80.4	0.44	81.0	82.7
5898_H01	0.15	78.5	88.8	0.43	78.6	82.0
6199_A03	0.24	72.4	77.9	0.41	72.4	76.1
6199_D06	0.23	75.8	79.2	0.41	75.8	81.1
6199_B05	0.20	80.2		0.43	78.4	
6199_B02	0.28	78.3	82.7	0.46	78.1	81.7
5898_E01				0.48	78.7	84.1
6199_E06	0.23	74.4	76.7	0.43	74.1	79.5
5898_F01	0.26	77.6	82.4	0.47	77.7	81.7
6199_D01	0.28	73.0	79.5	0.46	73.3	78.4
6199_D08	0.25	75.4	79.5	0.37	75.5	77.7
6199_E03	0.29	74.5	79.9	0.48	74.4	76.8
5898_C02	0.25	76.8	78.5	0.42	76.7	83.2
6199_F01	0.25	82.2	85.2	0.37	81.9	86.0
6199_D07	0.11	79.1	82.6	0.45	79.1	80.4
6199_C05	0.25	75.6		0.41	74.7	
6199_B03	0.27	78.4	80.9	0.48	78.0	81.6
6199_F08	0.21	78.6	87.7	0.45	78.6	83.1
6199_A05	0.15	77.0		0.47	76.3	
6199_G07	0.27	71.7		0.47	73.4	
6199_A07	0.12	75.5		0.43	75.5	

References

[1] C. Levinthal. How to fold graciously. In *Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois*, pages 22–24. University of Illinois Press, 1969.

[2] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins: Structure, Function, and Bioinformatics*, 21(3):167–195, 1995.

[3] P. E. Leopold, M. Montal, and J. N. Onuchic. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proceedings of the National Academy of Sciences of the United States of America*, 89(18):8721–8725, 1992.

[4] A. N. Fedorov and T. O. Baldwin. Cotranslational protein folding. *Journal of Biological Chemistry*, 272(52):32715–32718, 1997.

[5] N. T. Southall, K. A. Dill, and A. Haymet. A view of the hydrophobic effect. *The Journal of Physical Chemistry B*, 106(3):521–533, 2002.

[6] M. Sadqi, L. J. Lapidus, and V. Muñoz. How fast is protein hydrophobic collapse? *Proceedings of the National Academy of Sciences of the United States of America*, 100(21):12117–12122, 2003.

[7] D. Nettels, I. V. Gopich, A. Hoffmann, and B. Schuler. Ultrafast dynamics of protein collapse from single-molecule photon statistics. *Proceedings of the National Academy of Sciences of the United States of America*, 104(8):2655–2660, 2007.

[8] R. Gilmanshin, S. Williams, R. H. Callender, W. H. Woodruff, and R. B. Dyer. Fast events in protein folding: relaxation dynamics of secondary and tertiary structure in native apomyoglobin. *Proceedings of the National Academy of Sciences of the United States of America*, 94(8):3709–3713, 1997.

[9] D. L. Smith, Y. Deng, and Z. Zhang. Probing the non-covalent structure of proteins by amide hydrogen exchange and mass spectrometry. *Journal of Mass Spectrometry*, 32(2):135–146, 1997.

[10] T. E. Wales and J. R. Engen. Hydrogen exchange mass spectrometry for the analysis of protein dynamics. *Mass Spectrometry Reviews*, 25(1):158–170, 2006.

[11] A. Mittermaier and L. E. Kay. New tools provide new insights in NMR studies of protein dynamics. *Science*, 312(5771):224–228, 2006.

[12] F. A. Mulder, A. Mittermaier, B. Hon, F. W. Dahlquist, and L. E. Kay. Studying excited states of proteins by NMR spectroscopy. *Nature Structural & Molecular Biology*, 8(11):932–935, 2001.

[13] A. J. Baldwin and L. E. Kay. NMR spectroscopy brings invisible protein states into focus. *Nature Chemical Biology*, 5(11):808–814, 2009.

[14] F. Chiti, M. Stefani, N. Taddei, G. Ramponi, and C. M. Dobson. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature*, 424(6950):805–808, 2003.

[15] S. N. Cohen, A. C. Chang, H. W. Boyer, and R. B. Helling. Construction of biologically functional bacterial plasmids in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, 70(11):3240–3244, 1973.

[16] F. Baneyx. Recombinant protein expression in Escherichia coli. *Current Opinion in Biotechnology*, 10(5):411–421, 1999.

[17] S. Y. Lee. High cell-density culture of Escherichia coli. *Trends in Biotechnology*, 14(3):98–105, 1996.

[18] F. Baneyx and M. Mujacic. Recombinant protein folding and misfolding in Escherichia coli. *Nature Biotechnology*, 22(11):1399–1408, 2004.

[19] S. M. Singh and A. K. Panda. Solubilization and refolding of bacterial inclusion body proteins. *Journal of Bioscience and Bioengineering*, 99(4):303–310, 2005.

[20] G. A. Bowden, A. M. Paredes, and G. Georgiou. Structure and morphology of protein inclusion bodies in Escherichia coli. *Nature Biotechnology*, 9(8):725–730, 1991.

[21] K. Oberg, B. A. Chrunyk, R. Wetzel, and A. L. Fink. Native-like secondary structure in interleukin-1ß. inclusion bodies by attenuated total reflectance FTIR. *Biochemistry*, 33(9):2628–2634, 1994.

[22] T. M. Przybycien, J. P. Dunn, P. Valax, and G. Georglou. Secondary structure characterization of ß-lactamase inclusion bodies. *Protein Engineering*, 7(1):131–136, 1994.

[23] D. Ami, A. Natalello, G. Taylor, G. Tonon, and S. M. Doglia. Structural analysis of protein inclusion bodies by fourier transform infrared microspectroscopy. *Biochimica et Biophysica Acta (BBA)-Proteins & Proteomics*, 1764(4):793–799, 2006.

[24] D. Ami, A. Natalello, P. Gatti-Lafranconi, M. Lotti, and S. M. Doglia. Kinetics of inclusion body formation studied in intact cells by FT-IR spectroscopy. *FEBS Letters*, 579(16):3433–3436, 2005.

[25] E. García-Fruitós, N. González-Montalbán, M. Morell, A. Vera, R. M. Ferraz, A. Arís, S. Ventura, and A. Villaverde. Aggregation as bacterial inclusion bodies does not imply inactivation of enzymes and fluorescent proteins. *Microbial Cell Factories*, 4(1):27, 2005.

[26] P. Smialowski, G. Doose, P. Torkler, S. Kaufmann, and D. Frishman. PROSO II–a new method for protein solubility prediction. *FEBS Journal*, 279(12):2192–2200, 2012.

[27] S. B. Zimmerman and S. O. Trach. Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of Escherichia coli. *Journal of Molecular Biology*, 222(3):599–620, 1991.

[28] R. J. Ellis. Macromolecular crowding: obvious but underappreciated. *Trends in Biochemical Sciences*, 26(10):597–604, 2001.

[29] S. O. Enfors. Control of in vivo proteolysis in the production of recombinant proteins. *Trends in Biotechnology*, 10:310–315, 1992.

[30] J. Corchero, R. Cubarsi, S. Enfors, and A. Villaverde. Limited in vivo proteolysis of aggregated proteins. *Biochemical and Biophysical Research Communications*, 237(2):325–330, 1997.

[31] M. Carrió and A. Villaverde. Role of molecular chaperones in inclusion body formation. *FEBS Letters*, 537(1):215–221, 2003.

[32] M. Martínez-Alonso, A. Vera, and A. Villaverde. Role of the chaperone DnaK in protein solubility and conformational quality in inclusion body-forming Escherichia coli cells. *FEMS Microbiology Letters*, 273(2):187–195, 2007.

[33] A. Rokney, M. Shagan, M. Kessel, Y. Smith, I. Rosenshine, and A. B. Oppenheim. E. coli transports aggregated proteins to the poles by a specific and energy-dependent process. *Journal of Molecular Biology*, 392(3):589–601, 2009.

[34] T. R. Jahn and S. E. Radford. Folding versus aggregation: Polypeptide conformations on competing pathways. *Archives of Biochemistry and Biophysics*, 469(1):100–117, 2008.

[35] T. Kiefhaber, R. Rudolph, H.-H. Kohler, and J. Buchner. Protein aggregation in vitro and in vivo: a quantitative model of the kinetic competition between folding and aggregation. *Nature Biotechnology*, 9(9):825–829, 1991.

[36] F. Chiti, N. Taddei, F. Baroni, C. Capanni, M. Stefani, G. Ramponi, and C. M. Dobson. Kinetic partitioning of protein folding and aggregation. *Nature Structural & Molecular Biology*, 9(2):137–143, 2002.

[37] A. L. Fink. Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Folding and Design*, 3(1):R9–R23, 1998.

[38] M. Morell, R. Bravo, A. Espargaró, X. Sisquella, F. X. Avilés, X. Fernàndez-Busquets, and S. Ventura. Inclusion bodies: specificity in their aggregation process and amyloid-like structure. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1783(10):1815–1825, 2008.

[39] L. Wang, S. K. Maji, M. R. Sawaya, D. Eisenberg, and R. Riek. Bacterial inclusion bodies contain amyloid-like structure. *PLoS Biology*, 6(8):e195, 2008.

[40] N. S. de Groot, R. Sabate, and S. Ventura. Amyloids in bacterial inclusion bodies. *Trends in Biochemical Sciences*, 34(8):408–416, 2009.

[41] R. S. Rajan, M. E. Illing, N. F. Bence, and R. R. Kopito. Specificity in intracellular protein aggregation and inclusion body formation. *Proceedings of the National Academy of Sciences of the United States of America*, 98(23):13060–13065, 2001.

[42] C. Chothia and J. Janin. Principles of protein-protein recognition. *Nature*, 256(5520):705–708, 1975.

[43] C. J. Tsai, S. L. Lin, H. J. Wolfson, and R. Nussinov. Studies of protein-protein interfaces: A statistical analysis of the hydrophobic effect. *Protein Science*, 6(1):53–64, 1997.

[44] T. A. Larsen, A. J. Olson, and D. S. Goodsell. Morphology of protein–protein interfaces. *Structure*, 6(4):421–427, 1998.

[45] S. Jones and J. M. Thornton. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 93(1):13–20, 1996.

[46] O. Carugo and P. Argos. Protein-protein crystal-packing contacts. *Protein Science*, 6(10):2261–2263, 1997.

[47] B. Kobe, G. Guncar, R. Buchholz, T. Huber, B. Maco, N. Cowieson, J. Martin, M. Marfori, and J. Forwood. Crystallography and protein-protein interactions: biological interfaces and crystal contacts. *Biochemical Society Transactions*, 36:1438–1441, 2008.

[48] S. Dasgupta, G. H. Iyer, S. H. Bryant, C. E. Lawrence, and J. A. Bell. Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers. *Proteins: Structure, Function, and Bioinformatics*, 28(4):494–514, 1997.

[49] M. J. Bennett, M. P. Schlunegger, and D. Eisenberg. 3D domain swapping: a mechanism for oligomer assembly. *Protein Science*, 4(12):2455–2468, 1995.

[50] M. Sunde, L. C. Serpell, M. Bartlam, P. E. Fraser, M. B. Pepys, and C. C. Blake. Common core structure of amyloid fibrils by synchrotron x-ray diffraction. *Journal of Molecular Biology*, 273(3):729–739, 1997.

[51] T. Shibui, M. Uchida-Kamizono, H. Okazaki, J. Kondo, S. Murayama, Y. Morimoto, K. Nagahari, and Y. Teranishi. High-level secretion of human apolipoprotein E produced in Escherichia coli: use of a secretion plasmid containing tandemly polymerized ompF-hybrid gene. *Journal of Biotechnology*, 17(2):109–120, 1991.

[52] D. E. McRee. *Practical Protein Crystallography*. Academic Press, 1999.

[53] S. Jevševar, V. Gaberc-Porekar, I. Fonda, B. Podobnik, J. Grdadolnik, and V. Menart. Production of nonclassical inclusion bodies from which correctly folded protein can be extracted. *Biotechnology Progress*, 21(2):632–639, 2005.

[54] Š. Peternel, J. Grdadolnik, V. Gaberc-Porekar, and R. Komel. Engineering inclusion bodies for non denaturing extraction of functional proteins. *Microbial Cell Factories*, 7(1):34, 2008.

[55] K. Tsumoto, M. Umetsu, I. Kumagai, D. Ejima, and T. Arakawa. Solubilization of active green fluorescent protein from insoluble particles by guanidine and arginine. *Biochemical and Biophysical Research Communications*, 312(4):1383–1386, 2003.

[56] K. Tsumoto, M. Umetsu, I. Kumagai, D. Ejima, J. S. Philo, and T. Arakawa. Role of arginine in protein refolding, solubilization, and purification. *Biotechnology Progress*, 20(5):1301–1308, 2004.

[57] M. Carrió, N. González-Montalbán, A. Vera, A. Villaverde, and S. Ventura. Amyloid-like properties of bacterial inclusion bodies. *Journal of Molecular Biology*, 347(5):1025–1037, 2005.

[58] C. P. Jaroniec, C. E. MacPhee, V. S. Bajaj, M. T. McMahon, C. M. Dobson, and R. G. Griffin. High-resolution molecular structure of a peptide in an amyloid fibril determined by magic angle spinning NMR spectroscopy. *Proceedings of the National Academy of Sciences of the United States of America*, 101(3):711–716, 2004.

[59] O. S. Makin, E. Atkins, P. Sikorski, J. Johansson, and L. C. Serpell. Molecular basis for amyloid fibril formation and stability. *Proceedings of the National Academy of Sciences of the United States of America*, 102(2):315–320, 2005.

[60] R. Nelson, M. R. Sawaya, M. Balbirnie, A. Ø. Madsen, C. Riekel, R. Grothe, and D. Eisenberg. Structure of the cross-β spine of amyloid-like fibrils. *Nature*, 435(7043):773–778, 2005.

[61] C. Wasmer, L. Benkemoun, R. Sabaté, M. O. Steinmetz, B. Coulary-Salin, L. Wang, R. Riek, S. J. Saupe, and B. H. Meier. Solid-state NMR spectroscopy reveals that E. coli inclusion bodies of HET-s (218–289) are amyloids. *Angewandte Chemie International Edition*, 48(26):4858–4860, 2009.

[62] W. E. Klunk, J. Pettegrew, and D. J. Abraham. Quantitative evaluation of congo red binding to amyloid-like proteins with a beta-pleated sheet conformation. *Journal of Histochemistry & Cytochemistry*, 37(8):1273–1281, 1989.

[63] H. Naiki, K. Higuchi, M. Hosokawa, and T. Takeda. Fluorometric determination of amyloid fibrils in vitro using the fluorescent dye, thioflavine T. *Analytical Biochemistry*, 177(2):244–249, 1989.

[64] H. Li, F. Rahimi, S. Sinha, P. Maiti, G. Bitan, and K. Murakami. Amyloids and protein aggregation - analytical methods. In R. Meyers, editor, *Encyclopedia of Analytical Chemistry: Supplementary Volume S1*, volume 1, pages 635–666. John Wiley & Sons, Ltd, 2011.

[65] L. Wang, D. Schubert, M. R. Sawaya, D. Eisenberg, and R. Riek. Multidimensional structure–activity relationship of a protein in its aggregated states. *Angewandte Chemie*, 122(23):3996–4000, 2010.

[66] L. Goldschmidt, P. K. Teng, R. Riek, and D. Eisenberg. Identifying the amylome, proteins capable of forming amyloid-like fibrils. *Proceedings of the National Academy of Sciences of the United States of America*, 107(8):3487–3492, 2010.

[67] F. Rousseau, J. Schymkowitz, and L. Serrano. Protein aggregation and amyloidosis: confusion of the kinds? *Current Opinion in Structural Biology*, 16(1):118–126, 2006.

[68] N. Gonzalez-Montalban, A. Natalello, E. García-Fruitós, A. Villaverde, and S. M. Doglia. In situ protein folding and activation in bacterial inclusion bodies. *Biotechnology and Bioengineering*, 100(4):797–802, 2008.

[69] E. García-Fruitós, A. Arís, and A. Villaverde. Localization of functional polypeptides in bacterial inclusion bodies. *Applied and Environmental Microbiology*, 73(1):289–294, 2007.

[70] N. S. De Groot and S. Ventura. Protein activity in bacterial inclusion bodies correlates with predicted aggregation rates. *Journal of Biotechnology*, 125(1):110, 2006.

[71] E. R. McCarney, J. E. Kohn, and K. W. Plaxco. Is there or isn't there? the case for (and against) residual structure in chemically denatured proteins. *Critical Reviews in Biochemistry and Molecular Biology*, 40(4):181–189, 2005.

[72] A. V. Finkelstein, A. Y. Badretdinov, and A. M. Gutin. Why do protein architectures have boltzmann-like statistics? *Proteins: Structure, Function, and Bioinformatics*, 23(2):142–150, 2004.

[73] E. E. Lattman, K. M. Fiebig, and K. A. Dill. Modeling compact denatured states of proteins. *Biochemistry*, 33(20):6158–6166, 1994.

[74] Y. J. Tan, M. Oliveberg, B. Davis, and A. R. Fersht. Perturbed pKa-values in the denatured states of proteins. *Journal of Molecular Biology*, 254(5):980–992, 1995.

[75] D. K. Wilkins, S. B. Grimshaw, V. Receveur, C. M. Dobson, J. A. Jones, and L. J. Smith. Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques. *Biochemistry*, 38(50):16424–16431, 1999.

[76] J. R. Gillespie and D. Shortle. Characterization of long-range structure in the denatured state of staphylococcal nuclease. I. paramagnetic relaxation enhancement by nitroxide spin labels. *Journal of Molecular Biology*, 268(1):158–169, 1997.

[77] J. R. Gillespie and D. Shortle. Characterization of long-range structure in the denatured state of staphylococcal nuclease. II. distance restraints from paramagnetic relaxation and calculation of an ensemble of structures. *Journal of Molecular Biology*, 268(1):170–184, 1997.

[78] T. Religa, J. Markson, U. Mayor, S. Freund, and A. Fersht. Solution structure of a protein denatured state and folding intermediate. *Nature*, 437(7061):1053–1056, 2005.

[79] J. Klein and P. Dhurjati. Protein aggregation kinetics in an Escherichia coli strain overexpressing a Salmonella typhimurium CheY mutant gene. *Applied and Environmental Microbiology*, 61(4):1220–1225, 1995.

[80] M. Vendruscolo and C. M. Dobson. Towards complete descriptions of the free–energy landscapes of proteins. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 363(1827):433–452, 2005.

[81] C. J. Tsai, S. Kumar, B. Ma, and R. Nussinov. Folding funnels, binding funnels, and protein function. *Protein Science*, 8(6):1181–1190, 1999.

[82] Y. Bai. Energy barriers, cooperativity, and hidden intermediates in the folding of small proteins. *Biochemical and Biophysical Research Communications*, 340(3):976–983, 2006.

[83] Y. Bai. Hidden intermediates and Levinthal paradox in the folding of small proteins. *Biochemical and Biophysical Research Communications*, 305(4):785–788, 2003.

[84] B. A. Chrunyk and R. Wetzel. Breakdown in the relationship between thermal and thermodynamic stability in an interleukin-1β point mutant modified in a surface loop. *Protein Engineering*, 6(7):733–738, 1993.

[85] R. Wetzel and B. A. Chrunyk. Inclusion body formation by interleukin-1β depends on the thermal sensitivity of a folding intermediate. *FEBS Letters*, 350(2):245–248, 1994.

[86] G. Marcon, G. Plakoutsi, and F. Chiti. Protein aggregation starting from the native globular state. *Methods in Enzymology*, 413:75–91, 2006.

[87] F. Bemporad and F. Chiti. Native-like aggregation of the acylphosphatase from Sulfolobus solfataricus and its biological implications. *FEBS Letters*, 583(16):2630–2638, 2009.

[88] F. Chiti and C. M. Dobson. Amyloid formation by globular proteins under native conditions. *Nature Chemical Biology*, 5(1):15–22, 2008.

[89] F. Rousseau, J. W. Schymkowitz, and L. S. Itzhaki. The unfolding story of three-dimensional domain swapping. *Structure*, 11(3):243–251, 2003.

[90] S. Sambashivan, Y. Liu, M. R. Sawaya, M. Gingery, and D. Eisenberg. Amyloid-like fibrils of ribonuclease a with three-dimensional domain-swapped and native-like structure. *Nature*, 437(7056):266–269, 2005.

[91] R. Pankov and K. M. Yamada. Fibronectin at a glance. *Journal of Cell Science*, 115(20):3861–3863, 2002.

[92] P. Bork and R. F. Doolittle. Proposed acquisition of an animal protein domain by bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 89(19):8990–8994, 1992.

[93] G. Baneyx and V. Vogel. Self-assembly of fibronectin into fibrillar networks underneath dipalmitoyl phosphatidylcholine monolayers: role of lipid matrix and tensile forces. *Proceedings of the National Academy of Sciences*, 96(22):12518–12523, 1999.

[94] T. Ohashi, D. P. Kiehart, and H. P. Erickson. Dynamics and elasticity of the fibronectin matrix in living cell culture visualized by fibronectin–green fluorescent protein. *Proceedings of the National Academy of Sciences of the United States of America*, 96(5):2153–2158, 1999.

[95] H. P. Erickson. Reversible unfolding of fibronectin type III and immunoglobulin domains provides the structural basis for stretch and elasticity of titin and fibronectin. *Proceedings of the National Academy of Sciences of the United States of America*, 91(21):10114–10118, 1994.

[96] G. Baneyx, L. Baugh, and V. Vogel. Fibronectin extension and unfolding within cell matrix fibrils controlled by cytoskeletal tension. *Proceedings of the National Academy of Sciences of the United States of America*, 99(8):5139–5143, 2002.

[97] D. C. Hocking, R. K. Smith, and P. J. McKeown-Longo. A novel role for the integrin-binding III-10 module in fibronectin matrix assembly. *The Journal of Cell BBiology*, 133(2):431–444, 1996.

[98] K. C. Ingham, S. A. Brew, S. Huff, and S. V. Litvinovich. Cryptic self-association sites in type III modules of fibronectin. *Journal of Biological Chemistry*, 272(3):1718–1724, 1997.

[99] C. Zhong, M. Chrzanowska-Wodnicka, J. Brown, A. Shaub, A. M. Belkin, and K. Burridge. Rho-mediated contractility exposes a cryptic site in fibronectin and induces fibronectin matrix assembly. *The Journal of Cell Biology*, 141(2):539–551, 1998.

[100] A. F. Oberhauser, C. Badilla-Fernandez, M. Carrion-Vazquez, and J. M. Fernandez. The mechanical hierarchies of fibronectin observed with single-molecule AFM. *Journal of Molecular Biology*, 319(2):433–447, 2002.

[101] E. P. Gee, D. Yüksel, C. M. Stultz, and D. E. Ingber. SLLISWD sequence in the 10FNIII domain initiates fibronectin fibrillogenesis. *Journal of Biological Chemistry*, 288(29):21329–21340, 2013.

[102] L. Li, H. H. L. Huang, C. L. Badilla, and J. M. Fernandez. Mechanical unfolding intermediates observed by single-molecule force spectroscopy in a fibronectin type III module. *Journal of Molecular Biology*, 345(4):817–826, 2005.

[103] E. Paci and M. Karplus. Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations. *Journal of Molecular Biology*, 288(3):441–459, 1999.

[104] M. Gao, D. Craig, V. Vogel, and K. Schulten. Identifying unfolding intermediates of FN-III 10 by steered molecular dynamics. *Journal of Molecular Biology*, 323(5):939–950, 2002.

[105] E. P. Gee, D. E. Ingber, and C. M. Stultz. Fibronectin unfolding revisited: modeling cell traction-mediated unfolding of the tenth type-III repeat. *PLoS One*, 3(6):e2373, 2008.

[106] R. Chiquet-Ehrismann. What distinguishes tenascin from fibronectin? *The FASEB Journal*, 4(9):2598–2604, 1990.

[107] S. P. Ng, R. W. Rounsevell, A. Steward, C. D. Geierhaas, P. M. Williams, E. Paci, and J. Clarke. Mechanical unfolding of TNfn3: the unfolding pathway of a fnIII domain probed by protein engineering, AFM and MD simulation. *Journal of Molecular Biology*, 350(4):776–789, 2005.

[108] M. Akke, J. Liu, J. Cavanagh, H. P. Erickson, and A. G. Palmer. Pervasive conformational fluctuations on microsecond time scales in a fibronectin type III domain. *Nature Structural & Molecular Biology*, 5(1):55–59, 1998.

[109] D. J. Leahy, W. A. Hendrickson, I. Aukhil, and H. P. Erickson. Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. *Science*, 258(5084):987–991, 1992.

[110] D. J. Leahy, I. Aukhil, and H. P. Erickson. 2.0 Å crystal structure of a four-domain segment of human fibronectin encompassing the RGD loop and synergy region. *Cell*, 84(1):155–164, 1996.

[111] E. Cota, S. J. Hamill, S. B. Fowler, and J. Clarke. Two proteins with the same structure respond very differently to mutation: the role of plasticity in protein stability. *Journal of Molecular Biology*, 302(3):713–725, 2000.

[112] E. Cota and J. Clarke. Folding of beta-sandwich proteins: Three-state transition of a fibronectin type III module. *Protein Science*, 9(1):112–120, 2000.

[113] E. Cota, A. Steward, S. B. Fowler, and J. Clarke. The folding nucleus of a fibronectin type III domain is composed of core residues of the immunoglobulin-like fold. *Journal of Molecular Biology*, 305(5):1185–1194, 2001.

[114] S. J. Hamill, A. Steward, and J. Clarke. The folding of an immunoglobulin-like greek key protein is defined by a common-core nucleus and regions constrained by topology. *Journal of Molecular Biology*, 297(1):165–178, 2000.

[115] D. Lipovšek. Adnectins: engineered target-binding protein therapeutics. *Protein Engineering Design and Selection*, 24(1-2):3–9, 2011.

[116] A. Koide, C. W. Bailey, X. Huang, and S. Koide. The fibronectin type III domain as a scaffold for novel binding proteins. *Journal of Molecular Biology*, 284(4):1141–1151, 1998.

[117] L. Xu, P. Aha, K. Gu, R. G. Kuimelis, M. Kurz, T. Lam, A. C. Lim, H. Liu, P. A. Lohse, L. Sun, S. Weng, R. W. Wagner, and D. Lipovšek. Directed evolution of high-affinity antibody mimics using mrna display. *Chemistry & Biology*, 9(8):933–942, 2002.

[118] R. Camphausen, D. Fabrizio, M. C. Wright, P. Gage, and J. Mendlein. Targeted therapeutics based on engineered proteins for tyrosine kinases receptors, including IGF-IR, November 21 2007. US Patent App. 12/312,725.

[119] R. Wetzel, L. J. Perry, and C. Veilleux. Mutations in human interferon gamma affecting inclusion body formation identified by a general immunochemical screen. *Nature Biotechnology*, 9(8):731–737, 1991.

[120] B. Chrunyk, J. Evans, J. Lillquist, P. Young, and R. Wetzel. Inclusion body formation and protein stability in sequence variants of interleukin-1 beta. *Journal of Biological Chemistry*, 268(24):18053–18061, 1993.

[121] M. Gosselin and D. Lipovšek. IGF1R-binding adnectins: Expression analysis and monomericity. Personal communication, 2013.

[122] LB (Luria-Bertani) liquid medium. *Cold Spring Harbor Protocols*, 2006(1):pdb.rec8141, 2006.

[123] K. F. Geoghegan, H. B. Dixon, P. J. Rosner, L. R. Hoth, A. J. Lanzetti, K. A. Borzilleri, E. S. Marr, L. H. Pezzullo, L. B. Martin, P. K. LeMotte, et al. Spontaneous α-N-6-phosphogluconoylation of a "His Tag" in Escherichia coli: The cause of extra mass of 258 or 178 Da in fusion proteins. *Analytical Biochemistry*, 267(1):169–184, 1999.

[124] J. C. Aon, R. J. Caimi, A. H. Taylor, Q. Lu, F. Oluboyede, J. Dally, M. D. Kessler, J. J. Kerrigan, T. S. Lewis, L. A. Wysocki, et al. Suppressing posttranslational gluconoylation of heterologous proteins by metabolic engineering of Escherichia coli. *Applied and Environmental Microbiology*, 74(4):950–958, 2008.

[125] U. Rinas and J. E. Bailey. Protein compositional analysis of inclusion bodies produced in recombinant Escherichia coli. *Applied Microbiology and Biotechnology*, 37(5):609–614, 1992.

[126] Y. S. Kim, J. S. Wall, J. Meyer, C. Murphy, T. W. Randolph, M. C. Manning, A. Solomon, and J. F. Carpenter. Thermodynamic modulation of light chain amyloid fibril formation. *Journal of Biological Chemistry*, 275(3):1570–1574, 2000.

[127] M. Ramírez-Alvarado, J. S. Merkel, and L. Regan. A systematic exploration of the influence of the protein stability on amyloid fibril formation in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, 97(16):8979–8984, 2000.

[128] S. Krishnan, E. Y. Chi, J. N. Webb, B. S. Chang, D. Shan, M. Goldenberg, M. C. Manning, T. W. Randolph, and J. F. Carpenter. Aggregation of granulocyte colony stimulating factor under physiological conditions: characterization and thermodynamic inhibition. *Biochemistry*, 41(20):6422–6431, 2002.

[129] C. Pace and J. Hermans. The stability of globular protein. *Critical Reviews in Biochemistry and Molecular Biology*, 3(1):1–43, 1975.

[130] A. Cooper, M. A. Nutley, and A. Wadood. Differential scanning microcalorimetry. In S. Harding and B. Chowdhry, editors, *Protein-Ligand Interactions: Hydrodynamics and Calorimetry*, pages 287–318. Oxford University Press, Oxford, NY, 2000.

[131] J. M. Sturtevant. Biochemical applications of differential scanning calorimetry. *Annual Review of Physical Chemistry*, 38(1):463–488, 1987.

[132] J. C. Wilks and J. L. Slonczewski. pH of the cytoplasm and periplasm of Escherichia coli: rapid measurement by green fluorescent protein fluorimetry. *Journal of Bacteriology*, 189(15):5601–5607, 2007.

[133] A. Koide, M. R. Jordan, S. R. Horner, V. Batori, and S. Koide. Stabilization of a fibronectin type III domain by the removal of unfavorable electrostatic interactions on the protein surface. *Biochemistry*, 40(34):10326–10333, 2001.

[134] D. L. Wilkinson and R. G. Harrison. Predicting the solubility of recombinant proteins in Escherichia coli. *Nature Biotechnology*, 9(5):443–448, 1991.

[135] C. N. Magnan, A. Randall, and P. Baldi. SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics*, 25(17):2200–2207, 2009.

[136] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[137] Q. Wang, J. L. Johnson, N. Y. Agar, and J. N. Agar. Protein aggregation and protein instability govern familial amyotrophic lateral sclerosis patient survival. *PLoS Biology*, 6(7):e170, 2008.

[138] T. E. Creighton. *Proteins: structures and molecular properties*. Macmillan, 1993.

[139] A. G. Street and S. L. Mayo. Intrinsic β-sheet propensities result from van der waals interactions between side chains and the local backbone. *Proceedings of the National Academy of Sciences of the United States of America*, 96(16):9074–9076, 1999.

[140] V. Muñoz and L. Serrano. Elucidating the folding problem of helical peptides using empirical parameters. *Nature Structural & Molecular Biology*, 1(6):399–409, 1994.

[141] M. J. Thompson, S. A. Sievers, J. Karanicolas, M. I. Ivanova, D. Baker, and D. Eisenberg. The 3D profile method for identifying fibril-forming segments of proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 103(11):4074–4078, 2006.

[142] A. M. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz, and L. Serrano. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature Biotechnology*, 22(10):1302–1306, 2004.

[143] S. Maurer-Stroh, M. Debulpaep, N. Kuemmerer, M. L. de la Paz, I. C. Martins, J. Reumers, K. L. Morris, A. Copland, L. Serpell, L. Serrano, J. W. H. Schymkowitz, and F. Rousseau. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nature Methods*, 7(3):237–242, 2010.

[144] I. Walsh, F. Seno, S. C. Tosatto, and A. Trovato. PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Research*, 42(W1):W301–W307, 2014.

[145] O. Conchillo-Solé, N. de Groot, F. Avilés, J. Vendrell, X. Daura, and S. Ventura. AGGRESCAN: a server for the prediction and evaluation of hot spots of aggregation in polypeptides. *BMC Bioinformatics*, 8(1):65, 2007.

[146] G. G. Tartaglia, A. P. Pawar, S. Campioni, C. M. Dobson, F. Chiti, and M. Vendruscolo. Prediction of aggregation-prone regions in structured proteins. *Journal of Molecular Biology*, 380(2):425–436, 2008.

[147] S. O. Garbuzynskiy, M. Y. Lobanov, and O. V. Galzitskaya. FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics*, 26(3):326–332, 2010.

[148] J. U. Bowie, R. Luthy, and D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253(5016):164–170, 1991.

[149] R. Lüthy, J. U. Bowie, and D. Eisenberg. Assessment of protein models with three-dimensional profiles. *Nature*, 356:83–85, 1992.

[150] B. Kuhlman and D. Baker. Native protein sequences are close to optimal for their structures. *Proceedings of the National Academy of Sciences of the United States of America*, 97(19):10383–10388, 2000.

[151] R. Das and D. Baker. Macromolecular modeling with Rosetta. *Annual Review of Biochemistry*, 77:363–382, 2008.

[152] N. S. de Groot, I. Pallarés, F. X. Avilés, J. Vendrell, and S. Ventura. Prediction of "hot spots" of aggregation in disease-linked polypeptides. *BMC Structural Biology*, 5(1):18, 2005.

[153] G. G. Tartaglia and M. Vendruscolo. The zyggregator method for predicting protein aggregation propensities. *Chemical Society Reviews*, 37(7):1395–1401, 2008.

[154] J. S. Richardson and D. C. Richardson. Natural β-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proceedings of the National Academy of Sciences of the United States of America*, 99(5):2754–2759, 2002.

[155] C. J. Epstein, R. F. Goldberger, and C. B. Anfinsen. The genetic control of tertiary protein structure: studies with model systems. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 28, pages 439–449. Cold Spring Harbor Laboratory Press, 1963.

[156] H. M. Berman, B. C. Narayanan, L. Di Costanzo, S. Dutta, S. Ghosh, B. P. Hudson, C. L. Lawson, E. Peisach, A. Prlić, P. W. Rose, et al. Trendspotting in the protein data bank. *FEBS letters*, 587(8):1036–1045, 2013.

[157] R. D. Schaeffer and V. Daggett. Protein folds and protein folding. *Protein Engineering Design and Selection*, page gzq096, 2010.

[158] K. D. Pruitt, T. Tatusova, G. R. Brown, and D. R. Maglott. NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Research*, 40(D1):D130–D135, 2012.

[159] M. Levitt. Nature of the protein universe. *Proceedings of the National Academy of Sciences of the United States of America*, 106(27):11079–11084, 2009.

[160] B. Rost. Twilight zone of protein sequence alignments. *Protein Engineering*, 12(2):85–94, 1999.

[161] A. L. Main, T. S. Harvey, M. Baron, J. Boyd, and I. D. Campbell. The three-dimensional structure of the tenth type III module of fibronectin: an insight into RGD-mediated interactions. *Cell*, 71(4):671–678, 1992.

[162] C. D. Dickinson, B. Veerapandian, X. P. Dai, R. C. Hamlin, N. H. Xuong, E. Ruoslahti, and K. R. Ely. Crystal structure of the tenth type III cell adhesion module of human fibronectin. *Journal of Molecular Biology*, 236(4):1079–1092, 1994.

[163] A. Fiser, R. K. G. Do, and A. Šali. Modeling of loops in protein structures. *Protein Science*, 9(9):1753–1773, 2000.

[164] E. Michalsky, A. Goede, and R. Preissner. Loops in proteins (LIP)a comprehensive loop database for homology modelling. *Protein Engineering*, 16(12):979–985, 2003.

[165] D. J. Mandell, E. A. Coutsias, and T. Kortemme. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nature Methods*, 6(8):551–552, 2009.

[166] G. D. Friedland and T. Kortemme. Designing ensembles in conformational and sequence space to characterize and engineer proteins. *Current Opinion in Structural Biology*, 20(3):377–384, 2010.

[167] G. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7(1):95–99, 1963.

[168] N. Tokuriki, F. Stricher, J. Schymkowitz, L. Serrano, and D. S. Tawfik. The stability effects of protein mutations appear to be universally distributed. *Journal of Molecular Biology*, 369(5):1318–1332, 2007.

[169] W. R. Pestman. *Mathematical statistics: an introduction*. Walter de Gruyter, 1998.