

Old Dominion University

ODU Digital Commons

Theses and Dissertations in Biomedical
Sciences

College of Sciences

Spring 2015

Investigation Into Protein Folding and Misfolding

Jason Charles Collins
Old Dominion University

Follow this and additional works at: https://digitalcommons.odu.edu/biomedicalsciences_etds

Recommended Citation

Collins, Jason C.. "Investigation Into Protein Folding and Misfolding" (2015). Doctor of Philosophy (PhD), Dissertation, , Old Dominion University, DOI: 10.25777/gzq8-8k18
https://digitalcommons.odu.edu/biomedicalsciences_etds/27

This Dissertation is brought to you for free and open access by the College of Sciences at ODU Digital Commons. It has been accepted for inclusion in Theses and Dissertations in Biomedical Sciences by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.

INVESTIGATION INTO PROTEIN FOLDING AND MISFOLDING

By

Jason Charles Collins
B.S. Biochemistry, August 2008, Old Dominion University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

BIOMEDICAL SCIENCES

OLD DOMINION UNIVERSITY
May 2015

Approved By:

Lesley H. Greene (Director)

Jingdong Mao (Member)

Christopher Osgood (Member)

Patricia Pleban (Member)

ABSTRACT

INVESTIGATIONS INTO PROTEIN FOLDING AND MISFOLDING

Jason Charles Collins
Old Dominion University, 2015
Director: Dr. Lesley H. Greene

Proteins are the fundamental building blocks of all living organisms. They are critical in the proper function of virtually all cellular processes and without them biological life would be impossible. Since their discovery, understanding the transition of DNA to functional protein has been separated into two distinct areas of research; 1) how are they synthesized from genetic material and 2) once formed how do they fold into their native tertiary and quaternary structures. Understanding how genetic material encodes the primary structure of amino acids was the basis of the first-half of the genetic code. This in part sparked a technological race to solve the sequence of whole genomes, leading to the human genome project and its completion in 2006. The second-half of the genetic code is known as the ‘protein folding problem,’ which is the primary focus of this work. This problem is focused on understanding the fundamental features that dictate how the primary structure of amino acids transition along a thermodynamic and kinetic pathway into a functionally folded native-state.

The protein folding problem is a multifaceted and interdisciplinary area of research. Two key avenues of investigation include the folding of a protein into its native structure and the second is misfolding into an amyloid fibril structure. In the first-half of this work, we investigated the folding of the B1 domain of the *Streptococcal* immunoglobulin-binding domain of protein G (GB1) as our model system. Using bioinformatics approaches we investigated the origin and evolution of GB1. We also

elucidated a group of conserved residues and a network of long-range interactions which we propose are key determinants in dictating the stability, folding and native structure of the protein. GB1 was initially characterized using a combination of biophysical and high-resolution techniques such as stopped-flow and equilibrium fluorescence, circular dichroism and nuclear magnetic resonance spectroscopy. A microbial system was developed to facilitate testing the role of conserved long-range interactions through site-specific ^{13}C -labeling of tryptophan and phenylalanine. In the second-half, we investigated the transition of the Fas-associated death domain, an all α -helical Greek-key protein, into a misfolded amyloid-like state using additional techniques such as transmission electron and atomic force microscopy. From this work we were able to determine the extreme non-physiological conditions that would be required to allow for this transition. This result supports the 'generic amyloid hypothesis' that proposes that all proteins have the ability to form this alternative structure. In summary, this body of research has contributed to further advancing our understanding of the protein folding problem and laid the foundation for future atomic-level resolution studies.

This dissertation is dedicated to my lord and savior Jesus Christ and my parents, Waltraud and William Collins Jr., who have continually and unrelentingly provide me with love and support particularly through the hardest times of my career. Without them I would not have made it through the difficulties associated with the completion of a Ph.D.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my mentor, Dr. Lesley H. Greene for her dedicated assistance and support throughout my Ph.D. study at Old Dominion University. She has given me this excellent opportunity to think creatively, explore science, design and achieve ground breaking and very unique scientific research. This dissertation would not have been possible if it wasn't for her unending faith in me and my ability to conduct cutting-edge research. I could not have been more privileged to have the opportunity to learn new scientific methods and thanks to her I was able to expand my mind to include interdisciplinary research at the forefront of the field today. It is because of her dedication and support that I have the confidence to move forward as a scientific researcher.

I would like to extend thanks to the members of my dissertation committee, Dr. Patricia Pleban, Dr. Christopher J. Osgood and Dr. Jingdong Mao for their time and efforts in supporting me, providing me with insights and knowledge that has helped me to succeed. In addition, I would like to acknowledge my current lab group members, Nardos Sori, and John T. Bedford, for their support and contribution to this dissertation. I would also like to thank Dr. Janet Moloney, Dr. Hai Li, Dr. Agatha Munyanyi, Brian Butler and John Avis, for their helpful discussions and contributions to my scientific career. I am also grateful to Dr. Wei Cao at Applied Research Center for expert help with the transmission electron microscopy imaging. I would also like to recognize Isaiah Ruhl, Wassim Obeid, Dr. Junyan Zhong and Dr. Gina Hoatson for their help with and knowledge in the area of solution-state and solid-state NMR. I would also like to thank all the ODU faculty members who I have come to know over the years who have helped me to develop as a student and individual.

This dissertation research of J.C.C. is supported in part by funding from an Old Dominion University Fellowship, a CIBA Summer Fellowship and the FASEB (MARC) Travel Awards.

NOMENCLATURE

f_U	Fraction unfolded
Θ_{MRW}	Mean residue weight ellipticity
Θ_{obs}	Observed ellipticity
ΔG	Change in Gibb's free energy
$^{\circ}C$	Degrees Celsius
μl	Microliter
1D	One-dimensional
2D	Two-dimensional
3D	Three-dimensional
ACBP	Acylcoenzyme A-binding protein
AFM	Atomic force microscopy
AM	Amplitude modulation
Apaf-1	Apoptotic protease activating factor-1
ATP	Adenosine triphosphate
Aux	Aromatic auxotrophic <i>E. coli</i>
Barstar	Ribonuclease barnase inhibitor protein
BBL	Dihydrolipoamide succinyltransferase
BSA	Bovine serum albumin
$\beta 2M$	$\beta 2$ -microglobulin
βME	β -mercaptoethanol
CARD	Caspase recruitment domain
CASP	Critical assessment of techniques for protein structure prediction

CATH	Class, Architecture, Topology and Homology
CENT	Conserved Expanded Native Topology
CD	Circular dichroism
cDNA	Clonal deoxyribonucleic acid
CHARMM	Chemistry at Harvard Macromolecular Mechanics
CI2	Chymotrypsin inhibitor 2
CP	Cross-polarization
CR	Congo red
CRPK	Cross-peak
D ₂ O	Deuterium Oxide
Da	Daltons
DARR	Dipolar-assisted rotational resonance
DD	Death domain
DED	Death effector domain
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleoside triphosphate
DTT	Dithiothreitol
Fadd-DD	Fas-associated death domain protein
FM	Frequency Modulation
FT-IR	Fourier transformed infrared resonance
GA	Albumin-binding domain of protein G (Phage selected domain-1)
GAG	Glycosaminoglycan
GB1	Immunoglobulin G-binding B1domain of protein G

Gnd-HCl	Guanidine hydrochloride
GROMOS	Groningen molecular simulation
Hsp	Heat shock proteins
HSQC	Heteronuclear-single quantum coherence
IDP	Intrinsically disordered proteins
IgG	Immunoglobulin G
IPTG	Isopropyl β -D-1-thiogalactopyranoside
kV	Kilovolts
LB	Luria Broth
LSP-C	Lung surfactant protein C
M	Molarity
MAS	Magic-angle spinning
MD	Molecular dynamics
MEGA	Molecular Evolutionary Genetics Analysis
MHz	Megahertz
MUSCLE	Multiple sequence comparison by log-expectation
MWM	Molecular weight marker
MWCO	Molecular weight cut-off
Ni-NTA	Nickel-nitrilotriacetic acid
NMR	Nuclear magnetic resonance
OD	Optical density
PCR	Polymerase chain reaction
PDB	Protein data bank

Pfam	Protein families database
PG	Propylene glycol
PSI-BLAST	Position-specific iterative basic local alignment search tool
RNA	Ribonucleic Acid
Rpm	Rotations per minute
SCOP	Structural classification of proteins
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SOD	Superoxide dismutase
ssNMR	Solid-state nuclear magnetic resonance
StDev	Standard deviation
STM	Scanning tunneling microscopy
TEM	Transmission electron microscopy
ThT	Thioflavin T
T_m	Temperature midpoint
TTR	Transthyretin
U	Units
UV	Ultraviolet
WT	Wild-type

TABLE OF CONTENTS

	<i>Page</i>
LIST OF TABLES	xiv
LIST OF FIGURES	xv
 Chapter	
 I. INTRODUCTION	
PROTEIN FOLDING PROBLEM	1
METHODS TO STUDY PROTEIN FOLDING: EQUILIBRIUM FLUORESCENCE.....	23
METHODS TO STUDY PROTEIN FOLDING: Φ -VALUE ANALYSIS	26
METHODS TO STUDY PROTEIN FOLDING: STOPPED-FLOW FLUORESCENCE.....	30
METHODS TO STUDY PROTEIN FOLDING: FAR- & NEAR-UV CIRCULAR DICHROISM.....	33
METHODS TO STUDY PROTEIN FOLDING: NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY	40
PROTEIN MISFOLDING PROBLEM	45
METHODS TO STUDY PROTEIN MISFOLDING: THIOFLAVIN-T AND CONGO RED BINDING FLUORESCENCE.....	70
METHODS TO STUDY PROTEIN MISFOLDING: TRANSMISSION ELECTRON MICROSCOPY	74
METHODS TO STUDY PROTEIN MISFOLDING: ATOMIC FORCE MICROSCOPY.....	79
HIGH RESOLUTION ANALYSIS OF α -HELICAL PROTEIN FIBRILS AND THEIR POLYMORPHISMS	82
RESEARCH GOALS	95
 II. BIOINFORMATIC ANALYSIS OF THE IMMUNOGLOBULIN-BINDING AND THE ALBUMIN-BINDING DOMAINS OF PROTEIN G: A LOOK INTO A POSSIBLE EVOLUTIONARY ANCESTOR	
OVERVIEW	97
RESULTS AND DISCUSSION	104
MATERIALS AND METHODS.....	143
 III. THE PROPOSED ROLE OF CONSERVED RESIDUES IN THE STABILITY, STRUCTURE AND FOLDING OF THE B1 DOMAIN OF PROTEIN G BY SITE- DIRECTED MUTAGENESIS	
OVERVIEW	146
RESULTS AND DISCUSSION	149
MATERIALS AND METHODS.....	172

	<i>Page</i>
IV. DEVELOPMENT OF A METHOD TO ELUCIDATE THE FORMATION OF A LONG-RANGE INTERACTION IN THE B1 DOMAIN OF PROTEIN G USING NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY	
OVERVIEW	181
RESULTS AND DISCUSSION	184
MATERIALS AND METHODS	211
V. BIOPHYSICAL ANALYSIS OF THE TRANSITION OF AN ALL α -HELICAL GREEK-KEY PROTEIN INTO AMYLOID FIBRILS COMPOSED OF β -SHEET STRUCTURE	
OVERVIEW	216
RESULTS AND DISCUSSION	220
MATERIALS AND METHODS	234
VI. CONCLUSIONS AND FUTURE WORK	238
REFERENCES	246
APPENDICES	
I. BIOINFORMATICS OF GA AND GB1 EVOLUTION	293
II. METHODS FOR CLONING, EXPRESSION AND PURIFICATION OF WT-GB1 AND VARIANTS	353
III. NMR METHOD TO STUDY LONG-RANGE INTERACTION IN GB1	354
IV. FIGURE REPRINT PERMISSIONS	358
VITA	359

LIST OF TABLES

Table	<i>Page</i>
1. Examples of Small α -helical Polypeptide Hormones Known to be Functional Amyloids.....	53
2. List of All α -Helical Proteins Known to Form Amyloid Fibrils Clinically, <i>in vivo</i> or <i>in vitro</i>	58
3. List of Random Small Proteins from the Protein Data Bank.....	107
4. PSI-BLAST Common Sequences between GB1 and GA.....	110
5. Motifs Found in GB1 and GA Related Proteins using Pfam Analysis	137
6. Proteins Selected for Structural Alignment with GB1	150
7. Sequence Identity Analysis of GB1 and 13 Structurally Aligned Proteins	151
8. Summary of All Conserved Positions from the Conservation Analysis.....	160
9. Summary of Stopped-Flow Kinetics Data for WT- and F52Y-GB1	171
10. Primers Designed for Mutational Analysis of GB1	175
11. Stopped-Flow Kinetics Data of WT- and Phe52Tyr-GB1 at High Concentrations...	197
12. Reagent List for ^{13}C -Phe/ ^{13}C -Trp labeled GB1 Expression in Auxo(DE3)	213

LIST OF FIGURES

Figure	Page
1. Formation and structure of peptide bonds	3
2. Common amino acids and their structure	4
3. Classification of protein tertiary structure	5
4. Models for protein folding	8
5. Energy landscape scheme of protein folding and aggregation	16
6. Aromatic amino acids and schematic of equilibrium fluorescence	24
7. Energy diagram for Φ -value analysis of protein folding	29
8. Schematic of a traditional stopped-flow instrument	31
9. Origin of the CD effect	35
10. Far-UV CD spectra associated with various types of secondary structure	36
11. The Near-UV CD spectrum for Type II Dehydroquinase from <i>Streptomyces coelicolor</i>	38
12. Schematic of an NMR instrument.....	41
13. Schematic of the 2D DARR method.....	43
14. Mechanism of protein amyloidogenesis	46
15. Representative structures of amyloid fibrils and their common X-ray diffraction pattern	49
16. Schematic that represents the possible polypeptide fates following translation.....	52
17. Molecular structure of α -helical proteins related to amyloidogenic diseases.....	57
18. Molecular structure of small hormone α -helical proteins.....	63
19. Molecular structure of α -helical proteins not related to disease	65
20. Chemical structure of two commonly used molecular probes for amyloid fibril detection.....	71

Figure	Page
21. Congo red apple-green birefringence for the rare laryngeal amyloidosis.....	72
22. Schematic of the column design of a standard high-resolution TEM.....	76
23. A fundamental schematic of the design of AFM and example image obtained by AM-AFM.....	81
24. Schematic of the proposed polymorphic aggregation mechanism of insulin	83
25. Imaging and modeling of a non-helical protein and associated amyloid fibril morphologies.....	84
26. Circular polymorphism of apolipoprotein A-I amyloid fibrils	86
27. AFM imaging of insulin fibrils and their associated polymorphisms.....	88
28. Amyloid fibril imaging of other α -helical proteins associated with a disease state or functional amyloid form	90
29. Amyloid-like fibril imaging of α -helical proteins not associated with a disease state or functional amyloid forms.....	94
30. Representative schematic of a superfamily.....	99
31. Images of the folds of GB1 and GA	102
32. Structural images of ligand bound GB1 and GA	103
33. Single amino acid mutations leading to fold switching	105
34. MUSCLE sequence alignment of GB1 and GA	106
35. MUSCLE sequence alignment of GB1 and GA with 7 small random proteins	108
36. MUSCLE alignment of common PSI-BLAST protein sequences.....	111
37. Condensed and truncated schematic of MUSCLE alignment of common PSI-BLAST protein sequences	117
38. Phylogenetic tree of related sequences to GA from PSI-BLAST	120
39. Phylogenetic tree of related sequences to GB1 from PSI-BLAST	121

Figure	Page
40. Phylogenetic tree of 12 mucus-binding proteins	124
41. Phylogenetic analysis of all the related proteins from GB1 and GA.....	126
42. Schematic of the computational analysis and corresponding results.....	127
43. MUSCLE alignment of all related proteins	129
44. Condensed GB1 and GA sequences from Figure 42	135
45. SWISS-MODEL of GA and GB1 evolutionarily related proteins.....	139
46. I-TASSER models of GA and GB1 evolutionarily related proteins.....	141
47. Structural alignment of 1PGB and 2PTL.....	151
48. Finalized structure-based sequence alignment.....	152
49. Amino acid position conservation analysis.....	154
50. Amino acid character conservation analysis.....	157
51. Position specific hydropathy analysis.....	159
52. Circular dichroism spectra of WT- and Phe52Tyr-GB1	162
53. Thermal denaturation of WT- and Phe52Tyr-GB1	163
54. Equilibrium folding monitored by intrinsic tryptophan fluorescence.....	165
55. Image of the Lysine 31 and Tryptophan 43 proposed cation- π interaction	166
56. Fraction of unfolded protein versus Gnd-HCl	167
57. Stopped-flow kinetics of WT- and Phe52Tyr-GB1	170
58. Schematic of the Shikimate pathway showing the synthesis of aromatic amino acids in <i>E. coli</i>	185
59. Schematic of the lac operon, T7 promoter and its regulation of recombinant gene expression in <i>E. coli</i>	186
60. Titration of tester phage with auxotrophic lysogens	187

Figure	Page
61. Aromatic knockout verification of cultures of auxotrophic <i>E. coli</i> transfected with Phe52Tyr-GB1	188
62. Backbone structure of GB1 with highlighted core aromatics and specific ¹³ C labeling of Phe and Trp.....	190
63. ¹³ C Labeled carbons of interest in the core structure of GB1	191
64. Representative schematic of anticipated cross-peak formation from interactions of ¹³ C-specifically labeled WT-GB1	192
65. Schematic outline of a traditional versus folding-freezing stopped-flow instrumentation	195
66. Stopped-flow fluorescence kinetics of WT- and Phe52Tyr-GB1 at high protein concentrations	196
67. 1D ¹³ C Solution-state NMR spectrum of ¹³ C-Phe/ ¹³ C-Trp labeled WT-GB1	199
68. 1D ¹³ C ssNMR CP-MAS spectrum of ¹³ C-Phe/ ¹³ C-Trp labeled WT-GB1	200
69. 1D ¹³ C ssNMR spectrum of uniformly labeled WT-GB1	201
70. 2D ¹³ C- ¹³ C-DARR experiment on ¹³ C-Phe/ ¹³ C-Trp labeled WT-GB1	204
71. 1D ¹ H NMR spectrum of ¹³ C-Phe/ ¹³ C-Trp labeled WT-GB1 Vs pH.....	206
72. 2D ¹ H- ¹³ C HSQC spectrum of ¹³ C-Phe/ ¹³ C-Trp labeled WT-GB1 with chemical shift assignments	207
73. Analysis of pH dependence on the core of GB1 using ¹ H- ¹³ C HSQC NMR	208
74. Molten globule analysis of WT-GB1 at pH 2.0 by CD	210
75. Fadd-DD structure and thermal denaturation assay	219
76. Fibril formation studies using ThT fluorescence and Far-UV CD	222
77. Amyloid fibril formation time course monitored by far-UV CD	225
78. AFM images of Fadd-DD amyloid fibril formation	227
79. TEM images of Fadd-DD fibrils.....	228

Figure	Page
80. Effects of agitation on Fadd-DD fibril growth	230
81. Representative conditions delineating pathways towards and away from amyloid-like fibril formation.....	231

CHAPTER I

INTRODUCTION

PROTEIN FOLDING PROBLEM

Protein folding is an interdisciplinary area of research that has gained significant attention from the scientific community [1]. The increase in interest in protein folding is mostly due to the growth in protein sequences and the overwhelming appeal of solving one of the most fundamental cellular processes. A key aspect of biological life is the ability to synthesize functional proteins from genetic information. Proteins play a pivotal role in the survival and function of all organisms by performing a myriad of biological processes, from cell division to apoptosis. Since the discovery of DNA and the subsequent solving of its structure, understanding how the primary sequence of amino acids is synthesized from genetic material was the basis of the first-half of the genetic code. In addition, the complexity and magnitude of genetic material sparked a technological race to solve the sequence of whole genomes, leading to the human genome project and its completion in 2006 [2-5]. Solving the first-half of the genetic code led to the advancement of biological techniques which enhanced the ability of researchers to investigate the second-half of the genetic code, centered on understanding how the primary sequence dictates the native conformation of proteins.

Currently, there are two approaches to solving the protein folding problem. The first is to use computational methods to predict the three-dimensional (3D) structure of a protein from the sequence of amino acids and the second is to understand the relationship

between the folding mechanism and the protein sequence. Since the discovery by Christian Anfinsen that proteins have the ability to spontaneously fold into its native-state structure understanding this process has been a foundational area of biochemical research [1]. The structure and function of a protein's native-state is determined by the sequence of amino acids [1, 2]. Understanding how the amino acid sequence encodes and directs the folding of a protein into its native-state is termed the 'protein folding problem'. It is also significantly important to understand protein folding, because many protein misfolding diseases arise from the improper folding of proteins into their native-state. The archetypal example of a misfolded protein is the variant Glu6Val, of hemoglobin which leads to sickle cell anemia. Understanding the underlying determinants of the protein folding problem will provide the tools for protein engineering and the design of protein therapies.

Proteins are polymeric in nature. They are composed of a string of amino acids linked together by a covalent bond between the carboxyl carbons of one amino acid to the amine nitrogen of the following amino acid in a condensation reaction (Figure 1). The sequence of amino acids in the polymer makes up the primary structure of a protein. There are twenty naturally occurring amino acids which gives rise to a large diversity of possible protein sequences (Figure 2). In addition, they can be further diversified through amino acid modifications. The secondary structure of proteins is characterized by specific local structures which include α -helices, β -sheets and β -turns stabilized by hydrogen bonding of the protein backbone. The tertiary structure is defined as the overall topological arrangement of the secondary structure in 3D space. This would also be the native-state structure for a globular single domain protein. There are three major

classifications of tertiary structure; all α -helical, all β -sheet and mixed α/β (Figure 3).

Forming and maintaining the tertiary structure stability is in large part due to cumulative hydrophobic and ionic interactions, hydrogen bonding and disulfide bonds. The quaternary structure refers to the packing of several individual protein chains together into a higher order protein structure. The formation of multimeric states (i.e. homo- or heterodimers) also increases functional diversity and increases the complexity of the protein folding problem. It's quite amazing that despite the large number of divergent and functionally diverse proteins, nature evolved proteins to share common topologies, such as the Greek-key or the TIM barrel topology.

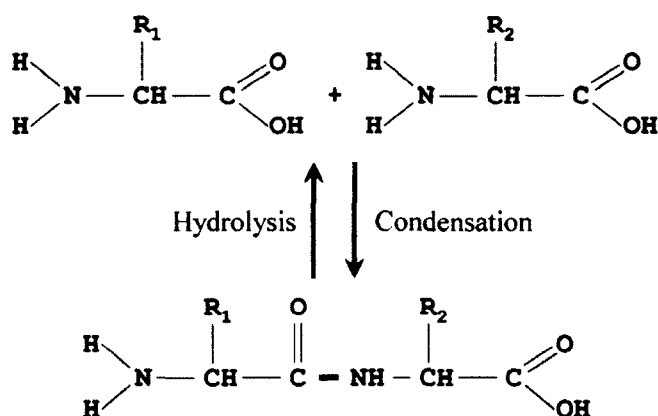


Figure 1. Formation and structure of peptide bonds. Peptide bonds can be formed and broken by condensation and hydrolysis reactions, respectively.

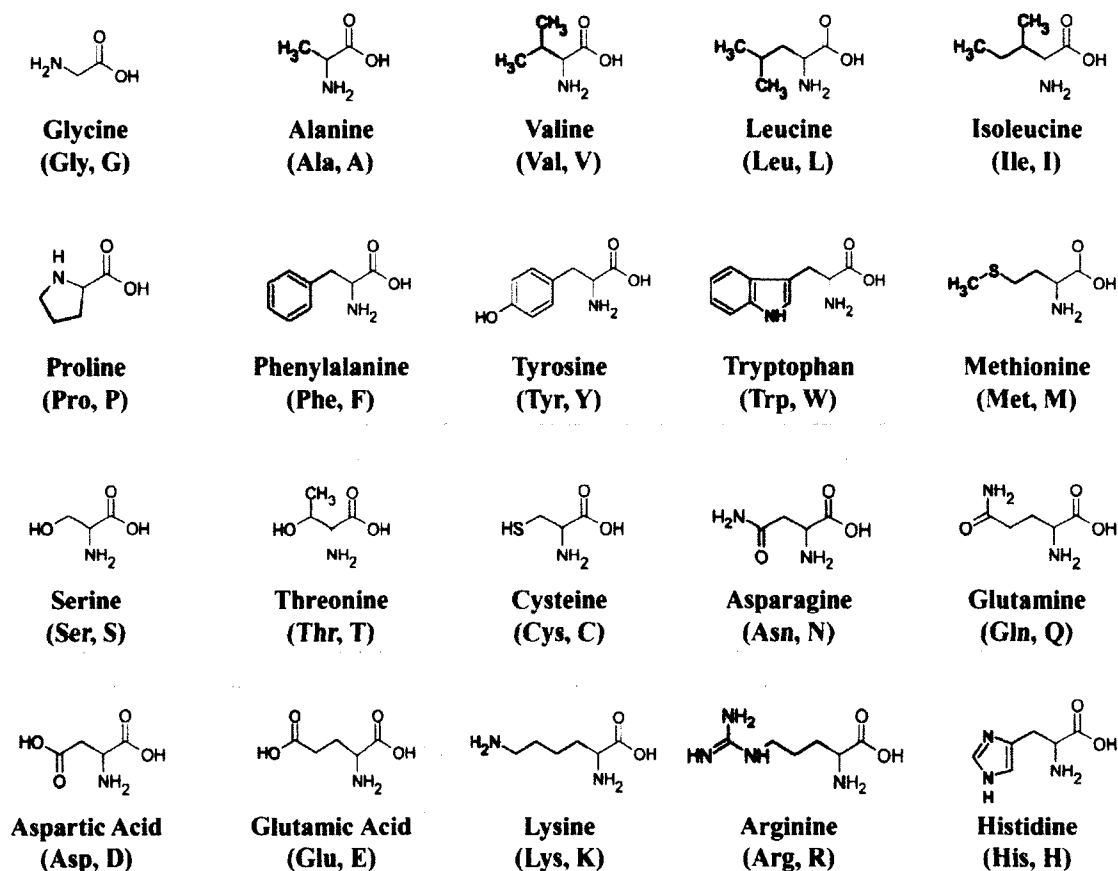


Figure 2. Common amino acids and their structure. Amino acids are chemically grouped by cyan borders into unsubstituted (dotted), aromatics (solid), hydroxyl (short-dash), carboxamide (long-dash), dicarboxylic (dash-dot-dot) and diamino (dash-dot) type. R-groups are highlighted and grouped by character: non-polar (blue), polar (brown), acidic (red) and basic (green). Proline, methionine and cysteine are grouped as heterocyclic, thioether and mercapto, respectively. These structures were drawn by hand in Microsoft Power Point (version 2010).

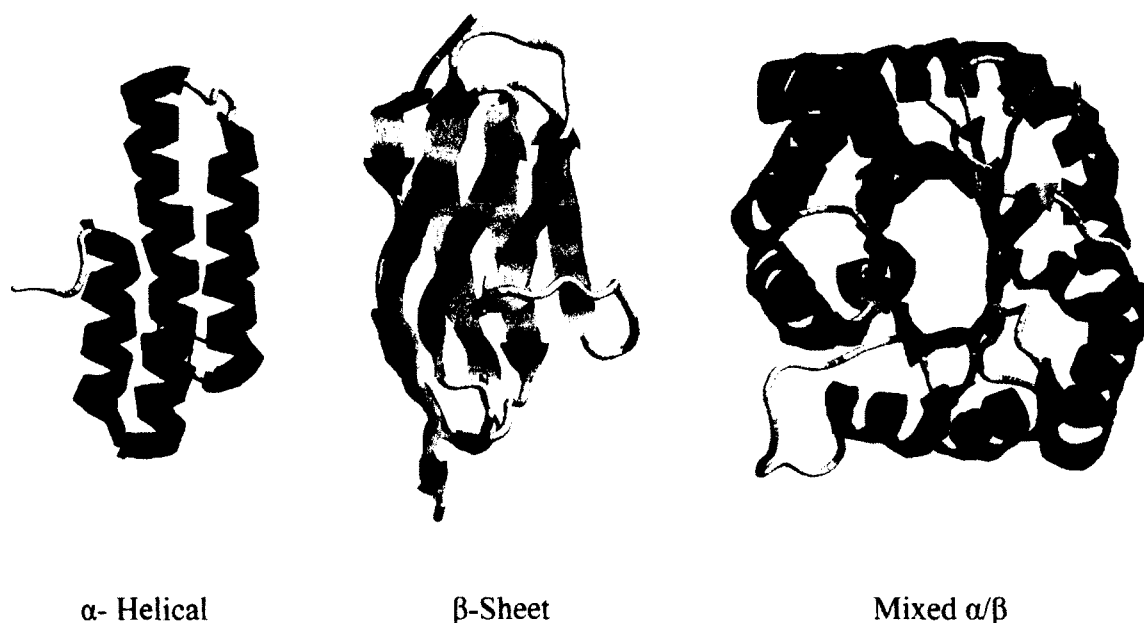


Figure 3. Classification of protein tertiary structure. Representations of each of the tertiary structure classes. The PDB codes for the structures shown are 1OP1 (Left), 1TIT (Center) and 1TIM (Right).

Cyrus Levinthal in the late 1960's postulated that if you had a protein whose primary structure contained 100 amino acids and folding occurred by attempting each possible conformation the length of folding time would be astronomical. Levinthal showed mathematically that even using a sampling rate of 10^{-4} nanoseconds this process would take 10^{27} years to find the native fold [6-8]. Since experimental evidence showed that proteins could fold on the millisecond time scale, Levinthal proposed that the process could not be random and the process must follow a specific pathway. Interestingly, directed folding was further supported by the proposal that sampling time could be significantly reduced if during the sampling process correct interactions were maintained while incorrect interactions are either not formed or not maintained and are broken to

continue sampling conformational space [7]. It was becoming clearer that protein folding was a directed process guided by underlying intermolecular interactions based on the work later done by Anfinsen. Anfinsen's dogma indicated that the well-defined native-state is predetermined solely on the interactions of the amino acids, which led to him winning the Nobel Prize in 1972 [9, 10].

Being able to visualize what a folding protein looks like and the orientation of each atom with angstrom-level resolution is vital to understanding how proteins fold. To solve the structure of a protein we use two primary methods, X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, both of which can solve the structure of a protein down to the atomic-level. Once protein structures are solved they are stored in the Protein Data Bank (PDB) for open access to the world. Being able to visualize these 3D forms allows us to classify proteins into groups dependent on topology. SCOP (Structural Classification of Proteins) and CATH (Class, Architecture, Topology and Homology) are the two major databases that are used to classify protein structures into related groups such as families and superfamilies [3, 4].

Proteins fold according to a pathway or mechanism that outlines the transition from the unfolded state to the native-state. There are currently four dominant mechanisms that describe the protein folding process: framework, hydrophobic collapse, nucleation-condensation and folding funnel models [11, 12]. It seems that in the universe of protein folding size is important in the folding of a protein. Small proteins generally fold rapidly in a two-state mechanism that involves only the denatured and native-states. The two-state mechanism of proteins containing less than 100 amino acids is characterized by

rapid folding that appear to have no kinetically detectable intermediates. However, larger proteins increase in complexity and are often described by a multiphasic process.

In the framework model, folding begins with the formation of secondary structure elements prior to the formation of the tertiary structure (Figure 4). Following the intermediate formation is the assembly of the secondary structure into a tightly packed native tertiary structure either by random assembly or by the careful propagation of ordered tertiary interactions [12, 13]. This model was highly supported by experimental work on small peptides early on and it was believed that the secondary structure could not form without some tertiary interactions due to lack in stability [8]. A few examples of proteins that demonstrates the framework model is acyl-coenzyme A-binding protein (ACBP), engrailed homeodomain and protein A which are all α -helical proteins and ribonuclease A, a mixed α/β protein [8, 14-16]. In the case of ACBP twenty-six residues were identified as evolutionarily conserved. Mutagenesis studies on ACBP showed that modification of conserved hydrophobic residues in helices 1 and 4 had a significant depressing effect on the folding rates [15]. In their investigation they determined that the formation of long-range interactions was favorable based on the spatial orientation of helix 1 and 4. This indicated that the short-range hydrogen bonding of both helix 1 and 4 formed prior to the formation of tertiary structure in order to bring certain hydrophobic residues into close proximity for initial tertiary structure formation. Once the initial stacking of helix 1 and 4 occurs, the final two helices form and collapse onto the structure forming the native topology. The formation of native interactions between two secondary structure elements as the rate-limiting step prior to tertiary structure formation in the folding of ACBP is an excellent example of a successive framework model [15]. In

addition, the framework model has been proposed as the mechanism for both T4 lysozyme and ribonuclease T1 based on high-resolution hydrogen-deuterium exchange experiments coupled with two-dimensional (2D) NMR [17, 18]. In both cases, the folding intermediates retained hydrogen bonds corresponding mostly to secondary structure features, indicating the formation of secondary structure prior to tertiary structure formation [17, 18].

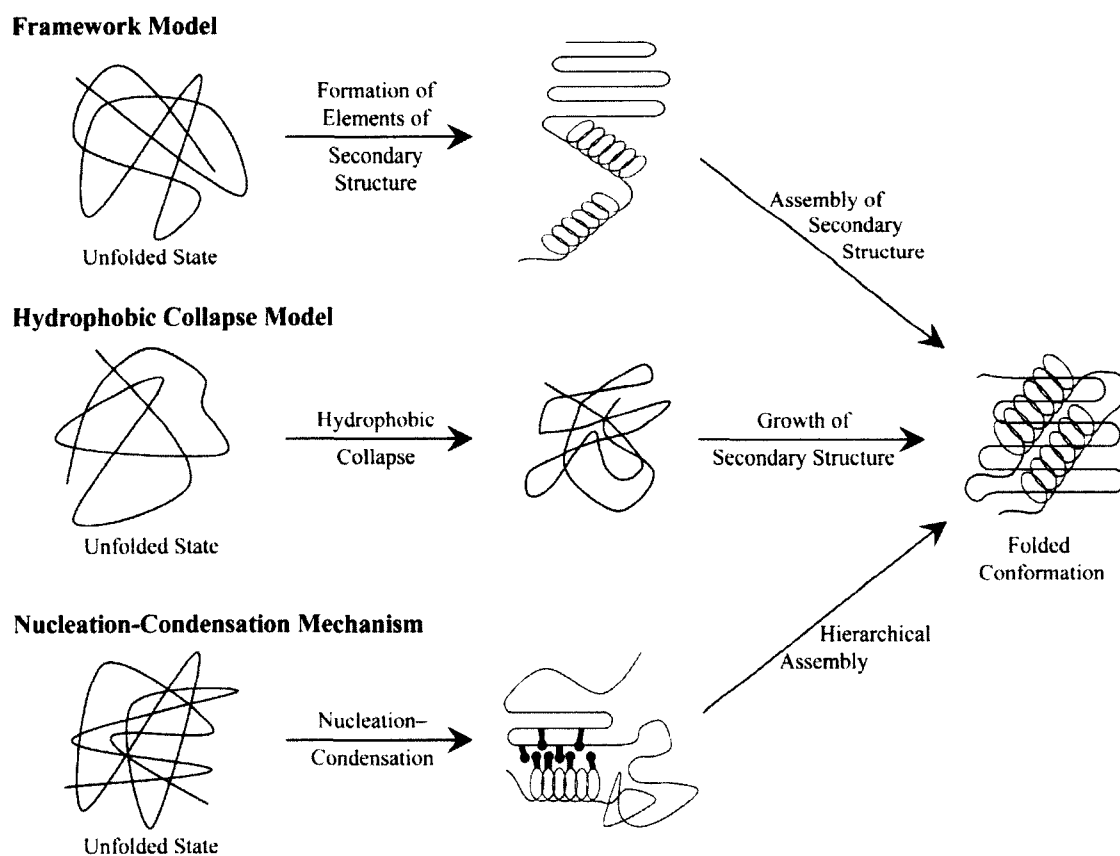


Figure 4. Models for protein folding. Shown is a schematic of the Framework model, Hydrophobic collapse and Nucleation-condensation mechanism models of protein folding. Image redrawn and modified from information and images in references [11, 12].

Interestingly, there is a predominance of α -helical type proteins that fold according to the framework model indicating that the formation of β -strands in this model appears to be relatively more difficult [8]. Similarly, the most commonly found secondary structures isolated in peptides was α -helical and turn structures and the least common β -hairpins [8, 19-21]. In this model the secondary structures formed from short peptide sequences could be considered starting points for protein folding. However, in practice these strong conformational preferences are not commonly seen and typically proteins will adopt a random coil for these peptides when included in a larger protein [22]. So it seems that the instability of these peptides in larger proteins is stabilized by the formation of tertiary interactions as seen in the investigation of a series of peptides of chymotrypsin inhibitor 2 (CI2) in which secondary structure did not form until the sequence of peptides included residues involved in long-range interactions [23].

In the hydrophobic collapse model of protein folding the unfolded protein initially collapses in a relatively uniform fashion, followed by a formation of ordered secondary structure elements (Figure 4) [24-26]. This process is directed by hydrophobic driving forces based on the expulsion of water by hydrophobic residues being attracted to each other in an aqueous environment [12]. It is believed that stable secondary structure can only form after the initial collapse of hydrophobic interactions. Folding of cytochrome *c*, an all α -helical protein containing a haem group, is one of the most well characterized proteins and is a relatively good model of the hydrophobic collapse [27]. The initial collapse of the unfolded cytochrome *c* involves the formation of an intermediate state which was shown by circular dichroism (CD) to contain about 20% of the native secondary structure content. In the fine-tuning phase continued compaction allows for

stabilizing contacts to form a second intermediate resembling a molten globule-like topology. The native conformation results from the folding of the remaining secondary elements and complete coordination to the haem group. In this case since most of the protein collapses into a hydrophobic core around the haem group prior to approximately 80% of the secondary structure forming, it seems that the hydrophobic collapse model better describes the folding mechanism of cytochrome *c* [27]. Interestingly, it seems that the haem group may function as a hydrophobic nucleus in which the initial collapse is focused. Another more controversial example of the hydrophobic collapse model is the folding of the ribonuclease barnase inhibitor protein (barstar).

The folding of barstar begins with a collapse of non-specific hydrophobic acids into a compact core intermediate without detectable secondary structure formation [28]. Rapid-mixing experiments on barstar in concert with secondary and tertiary structure detection via CD indicated that the biphasic rate constants for the formation of tertiary structure is 7000 s^{-1} and 11 s^{-1} whereas for secondary structure they are 4000 s^{-1} and 4 s^{-1} for the fast and slow phases, respectively [28]. The biphasic amplitude distributions for the rates of tertiary structure formation are 30 and 70 % whereas for the secondary structure formation they are both 50 %. The ANS binding experiment shows that 70 % of the hydrophobic collapse occurs in the 4 ms dead time prior to any observable folding. According to the data the rate of formation of the core is an order of magnitude greater than that for the formation of secondary or tertiary structure as calculated by CD. This data suggests that the folding follows the hydrophobic collapse model. However, a few years later it was shown experimentally using Φ -value analysis, which will be discussed later in this chapter, that the folding of barstar also follows the nucleation-condensation

model of protein folding [12, 29]. This was one of the first examples of a protein containing evidence of two folding models, which indicated that there could be a more unified mechanism of protein folding. In this example the phi-value analysis provided greater resolution into secondary and tertiary residue interactions that could not be discerned by CD.

The concept that folding occurs by a collapse of non-specific hydrophobic interactions has a few known intrinsic problems associated with it. One being that the excess interactions that could form in the tertiary structure during the collapse will limit the ability of the protein to reorganize the backbone and fine-tune amino acid orientations for the growth of the secondary structure elements [8]. This collapse would induce an energy barrier associated with dissolving and rearrangement of these interactions, which would be energetically unfavorable. Second, the denatured state may not be completely random and unstructured without any side-chain interactions. Instead, the unfolded protein is dynamic because there appears to be fluctuating interactions that can form variable secondary structures or even pockets of hydrophobic clusters [8]. Because the unfolded state may contain residual structure, which can be native-like or non-native, it is proposed that the initial hydrophobic collapse also results in the formation of some secondary structure elements [8]. It's unclear if the residual structure in the unfolded state is simply a limiting of solvent accessibility or if the interactions are important in the folding process by directing folding towards the native conformation. Thus, more than likely there is some specificity or organization to the hydrophobic collapse.

With the advances of experimental techniques for monitoring the fine details involved in the structural changes during the protein folding process it has become

clearer that the vast majority of proteins fold via the nucleation-condensation model (Figure 4) [8, 12, 30-33]. The development of the nucleation-condensation mechanism of folding came from two major events in science; the first being the recognition that proteins could fold in a simple two-state reaction mechanism without the accumulation of detectable intermediates and second is the development of Φ -value analysis technique which is an indirect method to detect the presence of interactions in the transition-state. This model for protein folding combines both the framework and hydrophobic collapse models in which some discrete secondary and tertiary contacts or regions form in a concerted process [8, 12, 33-35]. The initial collapse in the nucleation-condensation model is the formation of what is known as the folding nucleus made up of secondary and tertiary interactions that form the initial native-like topology and catalyze the folding process. These interactions and participating residues are now thought to be conserved in evolution. The folding nucleus is comprised of some native-like secondary interactions that are only stable in the presence of correctly formed long-range tertiary interactions [11, 12]. Following the collapse is the hierarchal assembly of native interactions, which are less conserved, as more structure condenses until the final stable native topology is achieved [36]. There are several well studied proteins like CI2, barstar, avian myeloblastosis transcription factor, tenascin, lysozyme and the 12-kDa FK506-binding protein that display the nucleation-condensation model of protein folding [11, 12, 29, 37-41].

The first example of this particular folding model is the folding of CI2 which is described as the archetypal protein. Investigations of scanning microcalorimetry and refolding kinetics showed that the folding of CI2 fits a two-state folding mechanism [34,

42, 43]. Detailed investigation of the role of specific residues on the transition-state of CI2 (a small protein consisting of 8 β -strands and 2 α -helices) using Φ -value analysis showed the presence of a folding nucleus and a stabilization of the helix by interactions between Ala16, Leu49 and Ile57 in the hydrophobic core. The initial nucleus had low stability which stabilizes as the protein forms more native-like contacts once the folding energy barrier has been overcome [43]. The folding of barstar, an 89-amino acid protein containing a three stranded parallel β -sheet and four α -helices, was shown to be a reversible two-state reaction mechanism. In the Φ -value analysis of barstar there is a stabilization of helix 1 and 4 in the initial consolidation towards the transition-state. The folding process for barstar is triphasic as the diffuse nucleus and its growth towards the transition state was mapped in three steps; 500 μ s, 1 ms and 100 ms. The native-like topology emerges after about 100 ms as the structure is stabilized by further propagation of tertiary and secondary interactions [12]. Experimental investigations of CI2 and barstar showed that they both fit the nucleation-condensation model of protein folding because they both fold in a two-state reaction mechanism and form a folding nucleus comprised of regions of secondary structure stabilized by a discrete set of correctly formed native tertiary interactions [12, 43]. The nucleation-condensation model may explain the rapid folding nature of proteins as opposed to a random sampling method. Interestingly, the expectations found in the folding funnel model, discussed in the next paragraph, coincide with data obtained from Φ -value analysis used to monitor the transition-state folding nucleus. The framework, hydrophobic and nucleation-condensation models are discussed in reference to small proteins with no intermediates and single transition-states. However,

each model can be related to larger proteins that may have several intermediate states during their folding reaction.

The folding funnel model begins with a large ensemble of unfolded conformations which transition down a funnel shaped energy landscape in which there is a decrease in the available conformational space during the folding reaction (Figure 5) [44, 45]. In this model depending on the environmental conditions the unfolding polypeptide can proceed towards a native-state topology or a misfolded state. If the environment is conducive to intramolecular contacts protein folding proceeds to the native-state, whereas if the environment is more favorable for intermolecular contacts protein misfolding can occur. The energy landscape model considers protein folding on a scale of free energy in which unfolded or partially folded intermediates of a higher free energy transition towards a global minimum free energy and entropy [46-49]. This model assumes that the native-state structure is considered the global free energy minimum. During the transition from the unfolded to the native fold, the conformational free energy decreases as the number of favorable native contacts increases. The protein folding funnel model begins with multiple folding pathways which later converge into a single pathway as more native contacts form. It is understood that hydrophobic residue sequestration away from the aqueous solvent is the underlying driving force in the folding funnel model to overcome the folding energy barrier. Similar to the multitude of unfolded conformations, the molten globule is believed to be an interrelated ensemble of folding intermediates that have formed from a hydrophobic collapse of fluctuating interactions. The *molten globule intermediate* is similar to the transition in the framework model containing most if not all of the secondary structures with very little tertiary

structure [50]. The molten globule may be present in all the previous models and some believe may belong to a unique class of proteins called intrinsically disorder structures [51]. The importance of the molten globule state is still under intense scrutiny, however its importance as an intermediate in the folding of several proteins, such as apomyoglobin, α -lactalbumin, calcium-binding lysozyme and cytochrome *c* is becoming more evident [52-59]. It is becoming increasingly more prevalent in protein research to isolate the molten globule state or key structural features *in vitro* under a specific set of conditions, such as low pH as it is easier to study an equilibrium form than a kinetic intermediate [57, 59-61].

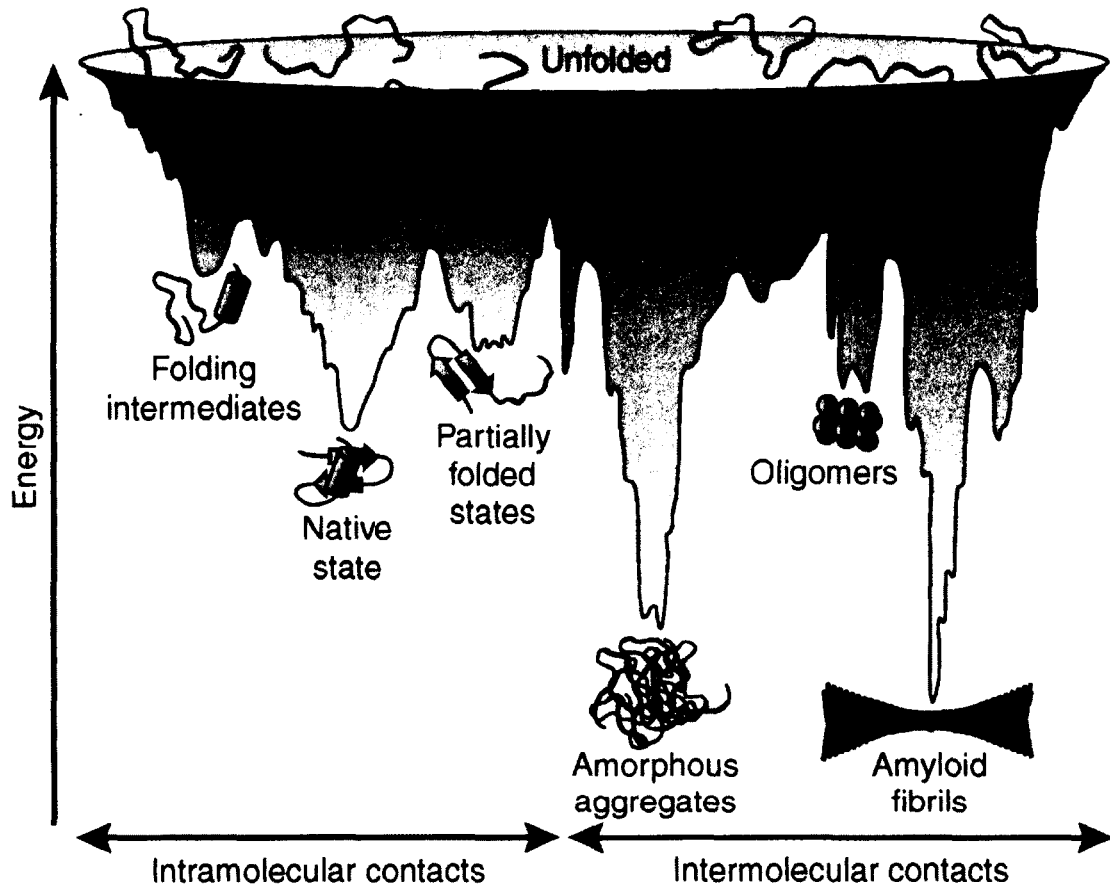


Figure 5. Energy landscape scheme of protein folding and aggregation. The blue surface shows the multitude of conformations ‘funneling’ to the native-state via intramolecular contacts and the purple area shows the conformations moving toward amorphous aggregates, off-pathway and potentially cell-toxic oligomers or low-energy amyloid fibrils via intermolecular contacts. This image is adapted from [62].

In relation to the folding funnel model the free energy of folding can be thermodynamically illustrated in the following equations:

$$\Delta G_{\text{total}} = (\Delta H_{\text{total}}) - T(\Delta S_{\text{total}}). \quad (1)$$

$$\Delta G_{\text{total}} = (\Delta H_{\text{protein}} + \Delta H_{\text{solvent}}) - T(\Delta S_{\text{protein}} + \Delta S_{\text{solvent}})_{\text{vent}}. \quad (2)$$

In the equations, ΔG_{total} is the change in total Gibbs free energy between the native and unfolded state of the folding reaction, which is related to the changes in enthalpy and entropy before and after the folding reaction [63]. ΔH_{total} is the sum of total enthalpy of the protein ($\Delta H_{\text{protein}}$) and the solvent ($\Delta H_{\text{solvent}}$). ΔS_{total} is the sum of total entropy of the protein ($\Delta S_{\text{protein}}$) and the solvent ($\Delta S_{\text{solvent}}$) [63]. The enthalpy and entropy of protein folding is directly related to the intermolecular forces and chemical characteristics of the amino acid groups that underlie protein folding. The stability of the protein is largely related to the contributions of the entropy and enthalpy of the solvent as it relates to the character of the amino acid side chains. Hydrophobic side chains provide a positive solvent entropy and the polar residues contribute a negative solvent enthalpy, which sum to an overall negative Gibbs free energy [63]. Since the free energy is negative this thoroughly supports the spontaneity of the protein folding process discovered by Anfinsen. The folding process and rate of the reaction is regulated by the rate at which the unfolded polypeptide chain overcomes a free energy barrier, catalyzed by the formation of a transition-state structure [63].

There are a multitude of theoretical studies that show that the protein folding rate is influenced by size, stability and topology [13, 30, 46, 64-81]. Interestingly, there is strong computational evidence for predicting the folding rates of proteins based on their relative contact order. Those proteins with greater local contacts tended to fold more rapidly than those with more non-local interactions [82]. The importance of local and non-local interactions in the native structure and their effect on the rate of the folding reaction was shown to significantly correlate to the relative contact order [82].

The mechanisms of protein folding have been previously described by experimental work, and the process has been shown to be spontaneous process *in vitro*. However, when we look at protein folding *in vivo* there are other factors that we need to account for and the process can be quite different. Proteins *in vivo* are only considered marginally stable which makes them highly susceptible to misfolding. In addition, the environment of the cell is quite different due to a high crowding effect (300-400 mg/ml of protein and other macromolecules) that is not seen *in vitro*, which can result in disruptions in the folding process leading to non-functional aggregates [63, 83, 84]. So far I have for the most part only discussed the folding mechanisms as they relate to small (~100 amino acids) monomeric single domain proteins with a simple two-state folding reaction. However, the complexity of protein folding increases astronomically as many proteins within the cell have multiple domains and the average protein size for eukaryotes is approximately 450 amino acids [85]. In addition, once folded many proteins gather into multimeric protein assemblies to perform their function, only further increasing the complexity. So how does the cell rapidly fold thousands of large proteins under what seems to be extremely unfavorable conditions as many proteins are intolerant to high concentrations? In addition, the range of available conditions (i.e. buffer, pH and temperature) for folding inside a cell are limited and very narrow unlike the wide range of *in vitro* conditions.

To compensate, the folding process can be assisted by molecular chaperones and chaperonin [84, 86]. The molecular chaperone network functions in a large array of proteostasis but one of its key functions is to prevent protein misfolding, which is particularly important during cellular stress. Any protein that interacts with and aids the

folding of another protein without being part of the final native structure defines the role of a chaperone. Chaperones interact with proteins either by direct interaction with the free nascent polypeptide chain or by interacting with the ribosome during translation and restricting unfavorable interactions [84]. Chaperonins and heat shock proteins (Hsps) are adenosine triphosphate (ATP) regulated chaperones which bind to the free nascent protein chain whereas, trigger factor, nascent-chain-associated complex and Hsp70L1 are some examples of ribosome bound chaperones that interact with the nascent protein chain as it exits the ribosomal tunnel [84]. In both cases, recognizing and binding hydrophobic segments of proteins that are typically found in the buried portion of the natively folded protein prevents the formation of improper intermolecular interactions [84]. Chaperones essentially act as kinetic partitions and redirect the polypeptide away from the pathway of aggregation by preventing the interaction of exposed hydrophobic residues. Binding the hydrophobic residues for a short time limits the available conformational space and allows formation of native interactions [84]. The importance of protein quality control chaperones cannot be over emphasized as there are many disorders related to deficiencies in proteostasis, such as the formation of aggregated α -synuclein sequestered into Lewy bodies in Parkinson's disease [87, 88].

One of the major goals in computational biochemistry has been solving the protein folding problem using computer based prediction methods. Using computational algorithms to fold the primary structure of amino acids into its native structure is fundamentally important. It has the potential to accelerate drug discovery by replacing lengthy structural studies with rapid structural simulations as well as provide insight into the structure and function of many as yet unsolved proteins in the genome [89]. There are

two major avenues in the approach to solving the structure of a protein using computational methods. The distinction in the two molecular dynamics (MD) approaches is that one relies on inference of structure and the other does not (*ab initio*). The more predominant approach to protein prediction uses protein databases to infer protein structure [89]. This originally began with databases for secondary structure prediction algorithms followed by the incorporation of computational physics such as atomic force fields and Monte Carlo simulation [90-94]. Further developments in the computational prediction of protein folding included homology modeling algorithms which looked at databases for homologous sequences to gain insight into structure [95]. Using these algorithms the fold of unknown sequences could be predicted based on a database of solved 3D structures with similar sequence and inferring similar folds [96]. In *ab initio* methods the protein structure is predicted only using known information about the underlying forces of protein folding.

The field of protein prediction was significantly advanced and influenced by the invention of the CASP (Critical Assessment of Techniques for Protein Structure Prediction) which was a community-wide test for the prediction of unknown structures [97]. Since the development of the CASP the native structure prediction of small, single-domain proteins using databases and advanced software have come to within 2-6 Å of the structure verified by experiment [98-102]. In addition, the advances in rapid homology modelling are impressively able to compute the approximate folds for whole genome sequences [103, 104]. Most significant and notable advances have occurred in the ability to align targets with homologs, detect evolutionarily distant homologs and the development of reasonable models for the most difficult proteins which excitingly

contain new folds [89, 98-100]. However, with these advances in native structure prediction there are still many obstacles to overcome. This includes refining the homology models, reducing the modeling error to consistently below 3 Å specifically for those large difficult proteins containing new folds, low homology or large β content [89, 98-100]. In addition, significant challenges still lie in the handling of large multidomain, domain-swapped proteins, membrane proteins and the prediction of protein-protein interactions [89, 98-100].

In the field of MD simulation for protein folding, one of the greatest challenges not related to protein folding, is to overcome available computing power. Although MD has been used to study the folding process from multiple aspects, the ultimate goal of MD is to develop an all-atom modelling method that encompasses the full folding process. However, even the fastest-folding proteins overwhelm the most advanced physics-based simulations when it comes to all-atom approaches requiring years of simulation time for even 1 millisecond folding time. It is quite impractical to run simulations for years, thus much of conventional MD is focused on specific characteristics of the folding process with little work being done on the complete folding trajectories of all atoms with atomic-resolution [105-107]. However, with the development of ANTON, a specialized supercomputer for high-speed MD simulations, the ability to compute all-atom, extended physics-based and explicit-solvent simulations has been extended by orders of magnitude in comparison to what was previously possible [108-110]. Using ANTON, detailed mechanistic analyses of the folding process for the folding of proteins up to ~100 amino acids which naturally fold rapidly can be examined [111, 112]. With the advent of ANTON it is estimated that 10% of the single-chain proteins in the protein data bank are

now accessible to MD simulations [108]. With this increase in accessibility of all-atom full folding pathway simulations, MD is now becoming more dependent on force field accuracy. There are a number of force field programs currently used in protein folding MD. A few of those most recently updated include Amber99SB-ILDN, GROMOS and CHARMM force fields [113-119]. These force-fields are shown to provide accurate representations of the structure and dynamics of small globular proteins on the submicrosecond scale, however they prove to be problematic on larger proteins due to longer folding timescales [120]. Now that MD simulation can exceed the millisecond timescale due to advanced computing power and the accuracy of the simulation is largely limited by the force field accuracy there has been a significant step in solving accurate fully atomistic physic-based models of fast-folding proteins [108].

METHODS TO STUDY PROTEIN FOLDING: EQUILIBRIUM

FLUORESCENCE

Studying protein folding, structure and stability in protein chemistry usually requires a spectroscopic probe that provides insight into the changing environment of the protein structure. Typical spectroscopic probes for protein structure and folding under equilibrium conditions utilize intrinsic fluorescence of the aromatic amino acids, mainly tryptophan (Figure 6A) [121-123]. However, tryptophan, phenylalanine and tyrosine typically exist in proteins naturally only at very low percentages i.e., approximately 1 mole percentage for tryptophan alone [124]. The application of tryptophan, phenylalanine and tyrosine relies on the maximal absorption wavelengths of their aromatic structure and natural hydrophobicity which makes them sensitive to the polarity of the immediate environment of each probe [124]. Phenylalanine, tyrosine and tryptophan absorb approximately 260 nm, 280 nm and 285 nm wavelengths of light respectively. Tryptophan is the most commonly used probe for investigating folding and unfolding due to its significantly larger molar extinction coefficient and fluorescence intensity in comparison to those of phenylalanine and tyrosine [124]. Changes in the local microenvironment of tryptophan typically correspond to an overall loss in native structure due to the cooperative nature of most unfolding transitions associated with many proteins [121-123]. Fluorescence probes allow for monitoring structural changes in a folding reaction induced by chemical denaturants, heat, pH or pressure [121]. In our investigation of the folding of the B1 immunoglobulin-binding domain of protein G (GB1) we used 295 nm as our excitation wavelength to specifically detect the intrinsic fluorescence of the single tryptophan at position 43. Typically a protein with a tryptophan

buried in its native structure will present with a shorter and higher maximum emission wavelength and intensity, respectively, in comparison to the denatured structure. Exposure of a tryptophan to a polar solvent reduces its emission potential as the excitation energy is dissipated by transfer into the solvent [124]. Consequently, the transition from native to denatured state results in a decrease in tryptophan fluorescence intensity (Figure 6B).

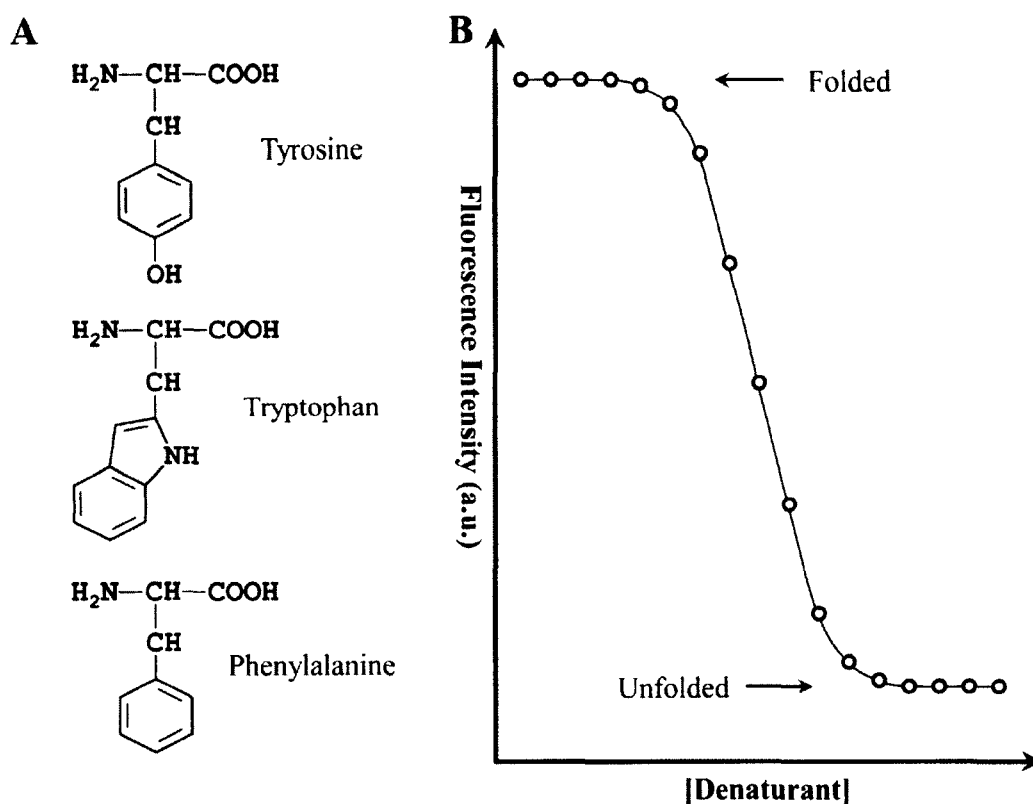


Figure 6. Aromatic amino acids and schematic of equilibrium fluorescence. (A)

Three aromatic residues used as intrinsic or synthetic spectroscopic probes of protein folding in equilibrium fluorescence experiments. **(B)** A visual schematic of equilibrium fluorescence data and the effect on aromatic residue emissions during the transition from unfolded to folded states.

Using the equilibrium fluorescence intensity in relationship to the denaturant is useful in determining free energy values of folding [121]. The inherent difficulties in using tryptophan and tyrosine residues as fluorophores for studying protein folding lie in the inability to distinguish the emission contribution of each individual amino acid when there are multiple residues in the primary structure [121, 125]. Proteins with multiple fluorophores are limited in interpretation of the structural and dynamic properties of the side chains due to overlap of emission signals in addition tryptophan fluorescence can be quenched by neighboring acid or basic residues, cysteine or disulfide bridges [121, 124, 125]. However, a potential solution to this limitation is to use site-directed mutagenesis to replace a specific amino acid, with non-fluorescent residue. In the case of a protein containing two fluorescent amino acids you can gain additional information by comparison of the single mutation to the wild-type (WT) protein which can reveal the importance of the specific amino acid in the stability and folding. An example of this has been shown using the tryptophan aporepressor protein from *E.coli* whose dimer interface appeared to be disrupted by the loss of either tryptophan [121]. Intrinsic tryptophan fluorescence has been used for many years to investigate protein folding and is foundational for chemical and thermal denaturation curves which provide perquisites for techniques such as Φ -value analysis and stopped-flow kinetic analysis.

METHODS TO STUDY PROTEIN FOLDING: Φ -VALUE ANALYSIS

Φ -value analysis is a protein engineering method originally initiated by the Fersht laboratory [126]. This method has seen much attention and has been used in the past two decades to investigate well over 500 variants of more than 21 different proteins [127]. Some examples of proteins that have been studied by Φ -value analysis include chymotrypsin inhibitor, SH3 domains, barnase, chicken brain α -spectin, ubiquitin, E3-binding domain of the dihydrolipoamide succinyltransferase (BBL), Fas-associated death domain (Fadd-DD) and individual domains of proteins L and G [128-135]. Using Φ -value analysis we can study the role of specific amino acids on the folding kinetics and conformational stability of the protein folding process by determining the ΔG of the transition-state relative to the folded state [136, 137]. This method of investigating the folding transition-state is most amenable to small domain proteins that follow a two-state transition reaction. The transition-state is relatively unstructured and because it is transient in solution, studying the structure of the transition-state is difficult for high resolution techniques like protein NMR or X-ray crystallography. Using site-directed mutagenesis select amino acids can be modified or removed by substitution with a similar amino acid or alanine, respectively. In the Φ -value analysis the effect of an engineered point mutation on the ΔG of the protein molecule is used as a probe of the structure formation along the folding pathway [126, 129, 136-138]. The Φ -value is obtained once the WT protein is compared with those of the engineered protein variants. The Φ -value is an indirect measure of the mutated residues molecular interaction contribution to the formation of the transition-state structure, and in turn reflects what percentage of the native-like interaction with the mutated residue remains in the transition-state [137].

Typically, positions that show the highest Φ -values are considered part of the folding nucleus as they are found in the most structured positions in the transition-state [137].

Due to advances in the modern laboratory techniques, mutations can be rapidly introduced into the plasmid construct using polymerase chain reaction (PCR) and readily purified using His-tagged or anion-exchange methods. In typical Φ -value analysis investigations, an impressive quantity of kinetic information can be gathered by rapidly mutating a large percentage of the available amino acids in a relatively short time. In this protein engineering method, selection of the appropriate mutations is quite important as non-conservative and disruptive mutations can result in a change in the native-state structure and alter the folding pathway completely [137]. In addition substituting a polar residue with a non-polar one can alter the proteins interaction with the solvent and result in solvation artifacts in the folding pathway [137]. Φ -value mutations typically are as conservative and mildly-disruptive as possible to prevent off-pathway folding, such as substituting Ile \rightarrow Val which removes a single methyl group. Although this minor change is quite sensible, typically researchers have opted to substitute alanine in any position selected for mutation for three main reasons [137]. First, using a unified substitution method is convenient, particularly when there are a high number of mutations that are being performed. Second, mutations that are too conservative often have changes that are too small and beyond the detectable limits of the method to calculate accurate Φ -values. Lastly, substitution of alanine of all the amino acids is the least likely to form new non-native interactions during the folding process and results in a sufficient loss of those interactions [137]. Interestingly, using Φ -value analysis on the substitution of a variety of amino acids at the same position can provide a significant amount of additional

information on the types of interactions that are important in transition-state stability [137]. The Φ_f value is defined using the following equations [137]:

$$\Phi_f = \Delta\Delta G_{t-u} / \Delta\Delta G_{f-u} = (\Delta G_{t-u}^{\text{WT}} - \Delta G_{t-u}^{\text{Mut}}) / (\Delta G_{f-u}^{\text{WT}} - \Delta G_{f-u}^{\text{Mut}}) \quad (3)$$

$$\Delta\Delta G_{f-u} = -RT \ln(\Delta k_f^{\text{Mut}} / \Delta k_f^{\text{WT}}) - RT \ln(\Delta k_u^{\text{Mut}} / \Delta k_u^{\text{WT}}) \quad (4)$$

$$\Delta\Delta G_{t-u} = -RT \ln(\Delta k_f^{\text{Mut}} / \Delta k_f^{\text{WT}}) \quad (5)$$

In the equation 3 $\Delta G_{t-u}^{\text{WT}}$ and $\Delta G_{t-u}^{\text{Mut}}$ represents the change in free energy between the transition-state and unfolding state of the WT and mutant respectively. $\Delta G_{f-u}^{\text{WT}}$ and $\Delta G_{f-u}^{\text{Mut}}$ represent the change in free energy between the folded and unfolded states of the WT and mutant respectively (Figure 7) [8, 137]. Thus, the Φ_f value is a ratio of the amount of intramolecular destabilization presented by the mutated residue in comparison to the native interaction from the WT protein. The Φ_f value ranges from 0 to 1 with values near 0 suggesting that the selected amino acid has no native interactions and is most likely surrounded by unstructured regions in the transitions state. Values near 1 indicate that the specific interactions of this particular residue position in the native-state are fully formed in the transition-state. Partial Φ_f values indicate that only a fraction of the native-state interactions are formed at that position in the transition-state. In Figure 7A, the effect of the mutation on the energy of the transition-state is equivalent to the native-state transition-state energy, which calculates to a $\Delta\Delta G_{t-u} = 0$, which results in a $\Phi_f = 0$. In opposition, in Figure 7B, the energy difference of the WT and mutant in the transition-states compared to the native-state results in $\Delta\Delta G_{t-u} = \Delta\Delta G_{f-u}$, which results in a $\Phi_f = 1$. $\Delta\Delta G_{t-u}$ and $\Delta\Delta G_{f-u}$ are calculated using equations 4 and 5 by substituting the folding rate constants (k_f) for the WT and mutants into the ratio. Experimentally, the Gibbs free energies for the WT and mutants are determined from chemical and thermal

denaturation folding curves. The folding rate constants and equilibrium constants, determined using stopped-flow measurements and a chevron plot, are required for calculation of the Φ_f [137].

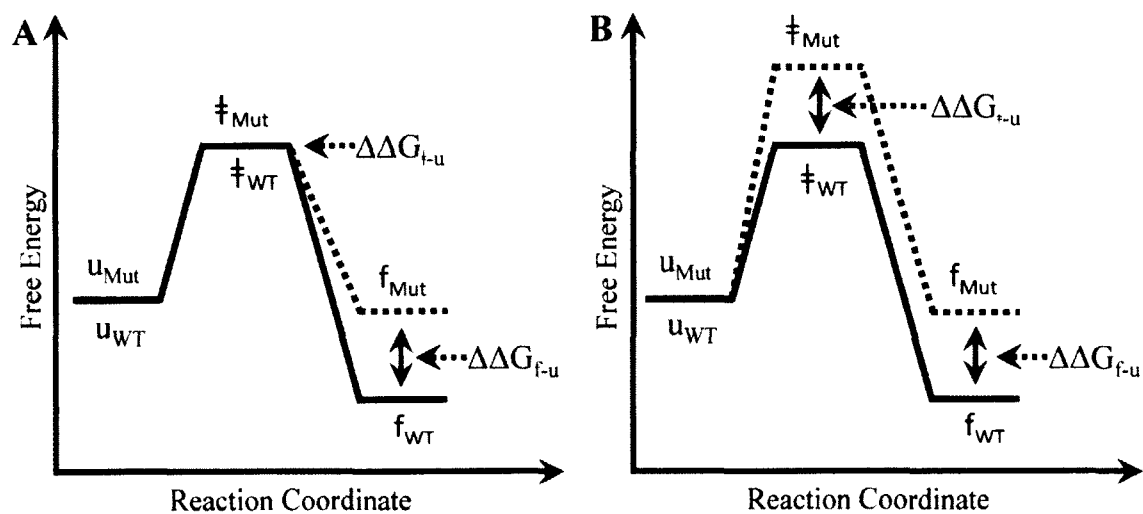


Figure 7. Energy diagram for Φ -value analysis of protein folding. The solid line shows energy levels of the unfolded, folded and transition-state for the WT protein, and the dotted line shows the levels for a mutant. (A) $\Phi = 0$ and (B) $\Phi = 1$. Modified and redrawn from references [8, 137].

METHODS TO STUDY PROTEIN FOLDING: STOPPED-FLOW SPECTROSCOPY

One of the most widely used methods for determining the folding kinetics of a protein is a rapid-mixing stopped-flow technique. In this type of experiment the folding process is initiated by rapidly mixing a chemically denatured protein into an appropriate refolding buffer initiating spontaneous refolding by dilution of the denaturant (Figure 8). In a similar fashion the unfolding of a protein can be monitored by rapidly mixing a native protein solution with a concentrated denaturant in an appropriate buffer. Forming denatured protein can be accomplished using several other methods in addition to chemicals, such as altering temperature and pH. The process of stopped-flow folding kinetics can be monitored in real-time by several spectroscopic methods including fluorescence, near- and far-UV CD, X-ray scattering, Fourier transform infrared spectroscopy (FT-IR) and real time NMR spectroscopy [50, 124, 139-144]. Using intrinsic fluorescence probes such as tryptophan emission radiation (excitation = 295 nm), the overall folding process can be monitored in real-time as the microenvironment of the tryptophan moves from a solvent exposed to a buried environment during the refolding process. Using CD, discussed in the next section, as the detection method allows for direct monitoring of structural information of a protein as it folds. These two detection methods are the most commonly used in the folding community [145]. Much of what is known about the kinetics of folding and unfolding is due to the enormous amount of information that has come from stopped-flow methods. There is a fundamental limitation to the information gathered by stopped-flow methods which is associated with poor time resolution. The time required to mix the solutions to initiate either unfolding or

refolding and move the solution into the observation head requires in general at least a few milliseconds. This time delay is called the “dead time” of the instrument. This puts a limit on the observable changes associated with folding, as in some cases all the spectroscopic changes have already occurred during the dead time [53, 146-148]. There are also studies that show that the formation of secondary structures such as α -helices in a synthetic polymer system is too fast to be observed by stopped-flow techniques [149]. However, there are some researchers dedicated to pushing the boundaries and have built systems that can monitor refolding on the microsecond time scale [150-152].

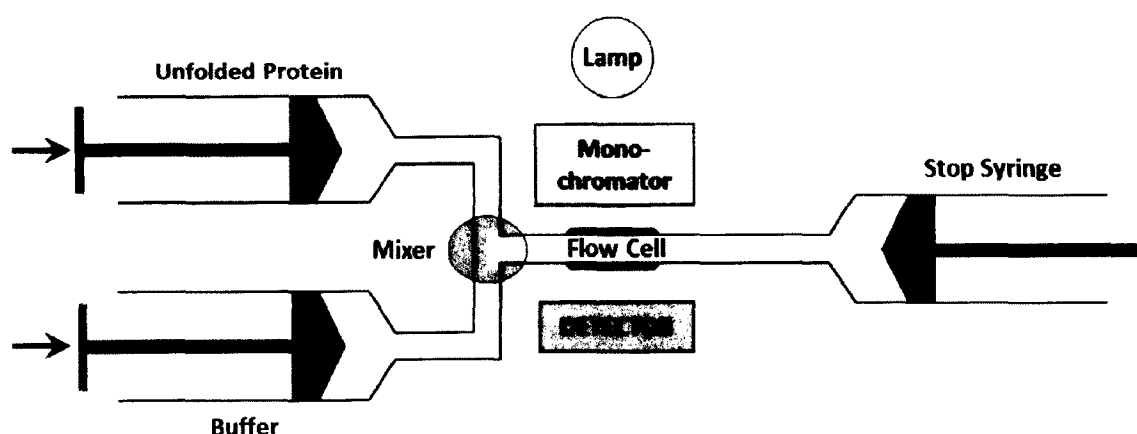


Figure 8. Schematic of a traditional stopped-flow instrument. Here the denatured protein in one syringe is rapidly mixed with refolding buffer from another syringe in the flow cell and monitored. The detector can be for example, fluorescence, absorption or CD. This figure was redrawn and adapted from [153].

In the past 20 years there have been a significant number of advances in our experimental and theoretical approaches to studying the mechanism of protein folding [47, 154-156]. Initially it was established that it was critically important to characterize the kinetics and thermodynamics of simple protein systems to establish a foundation to studying protein folding. It makes sense that the initial studies of protein folding would begin with small single domain proteins such as the 83 amino acid CI2, which does not contain any of the complexity of disulfides, cofactors or metal associations. Investigation of CI2 revealed a two-state folding mechanism with no kinetically detectable intermediates [34]. In following work on CI2, mutations were made and the relative effect on the folding rate and the participation of specific amino acids in the transition-state was monitored [31, 136]. The work with CI2 folding led to further investigations into the thermodynamics and kinetics of a number of other single domain proteins [1, 157]. Kinetics data obtained from stopped-flow methods are used to calculate folding rates and goodness of fit with multi-parameter exponentials using a graphing and fitting program like SigmaPlot.

METHODS TO STUDY PROTEIN FOLDING: FAR- & NEAR-UV CIRCULAR DICHROISM

In structural biology, CD is one of the fundamental spectroscopic methods for gathering structural information on proteins in biologically relatable environments. CD is a spectroscopic technique used in protein structure analysis because it can be used to identify the secondary and tertiary structure of proteins. This technique uses the differential absorption of circularly polarized UV light by chiral molecules (Figure 9A) [139, 158]. Unlike absorbance spectroscopy the left- and right-handed circular components of polarized light are absorbed differently by each of the chiral centers resulting in the CD spectrum (Figure 9B). In addition, non-protein cofactors like haem groups, pyridoxal-5'-phosphates, flavins and chlorophyll moieties can be seen in CD spectra [139, 159]. UV light in the far region (240 nm and below) of the spectrum are preferentially absorbed by the backbone amide of the peptide bond due to $n \rightarrow \pi^*$ and $\pi \rightarrow \pi^*$ electronic transitions, thus data obtained in this region correspond to the orientation of the polypeptide backbone or protein secondary structure (Figure 10) [139, 158, 159]. Far-UV CD allows for the determination of the three common types of secondary structures as well as disordered or irregular structure associated with denatured or intrinsically disordered proteins. For example, all α -helical proteins result in a strong negative differential of far-UV absorption with a distinctive double minimum around 208 and 222 nm. In contrast to that, all β -sheet proteins show a single minimum between 210 and 225 nm with a significant decrease in the strength of the negative differential [158]. Disordered structures show a loss of all peaks with a near 0 differential absorption between 205 and 250 nm. Advances in software analysis algorithms of far-UV spectra

allow for determination of the relative percentage each secondary structure contributes to the native structure of the protein molecule [139, 158]. This observation of the secondary structure is extremely useful in detecting changes in a protein structure due to changes in environmental conditions or functional changes. It is also quite valuable when predicting the structure of a protein, when compared to reference spectra of proteins whose structure has already been solved by either NMR or X-ray crystallography. There is a large collection of algorithms and datasets that are currently used to compare secondary structure. DICHROWEB is an online CD server that allows analysis of secondary composition directly from instrumental data, hosting the most commonly used algorithms including CDSSTR, SELCON, VARSLC, K2D and CONTIN [160-168]. However, there are limitations to the database analysis when it comes to small oligopeptide structures, which are not included in the database unless they are under conditions ideal for specific secondary structures.

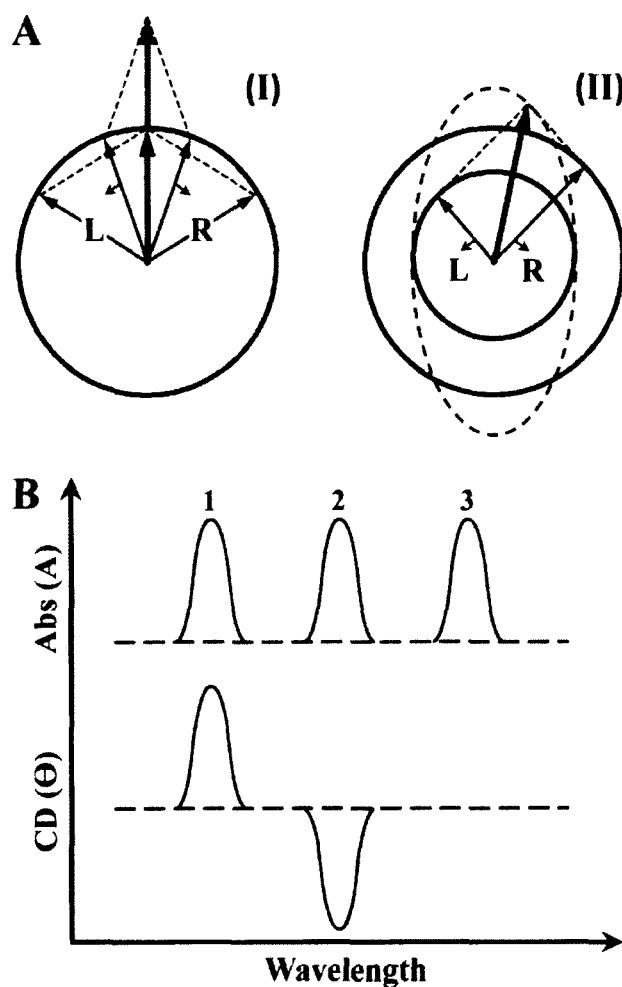


Figure 9. Origin of the CD effect. (A) The left (L) and right (R) circularly polarized components of plane polarized radiation: (I) the two components have the same amplitude and when combined generate plane polarized radiation; (II) the components are of different magnitude and the resultant (dashed line) is elliptically polarized. (B) The relationship between absorption and CD spectra. Band 1 has a positive CD spectrum with L absorbed more than R; band 2 has a negative CD spectrum with R absorbed more than L; band 3 is due to an achiral chromophore. This image was redrawn and modified from [139].

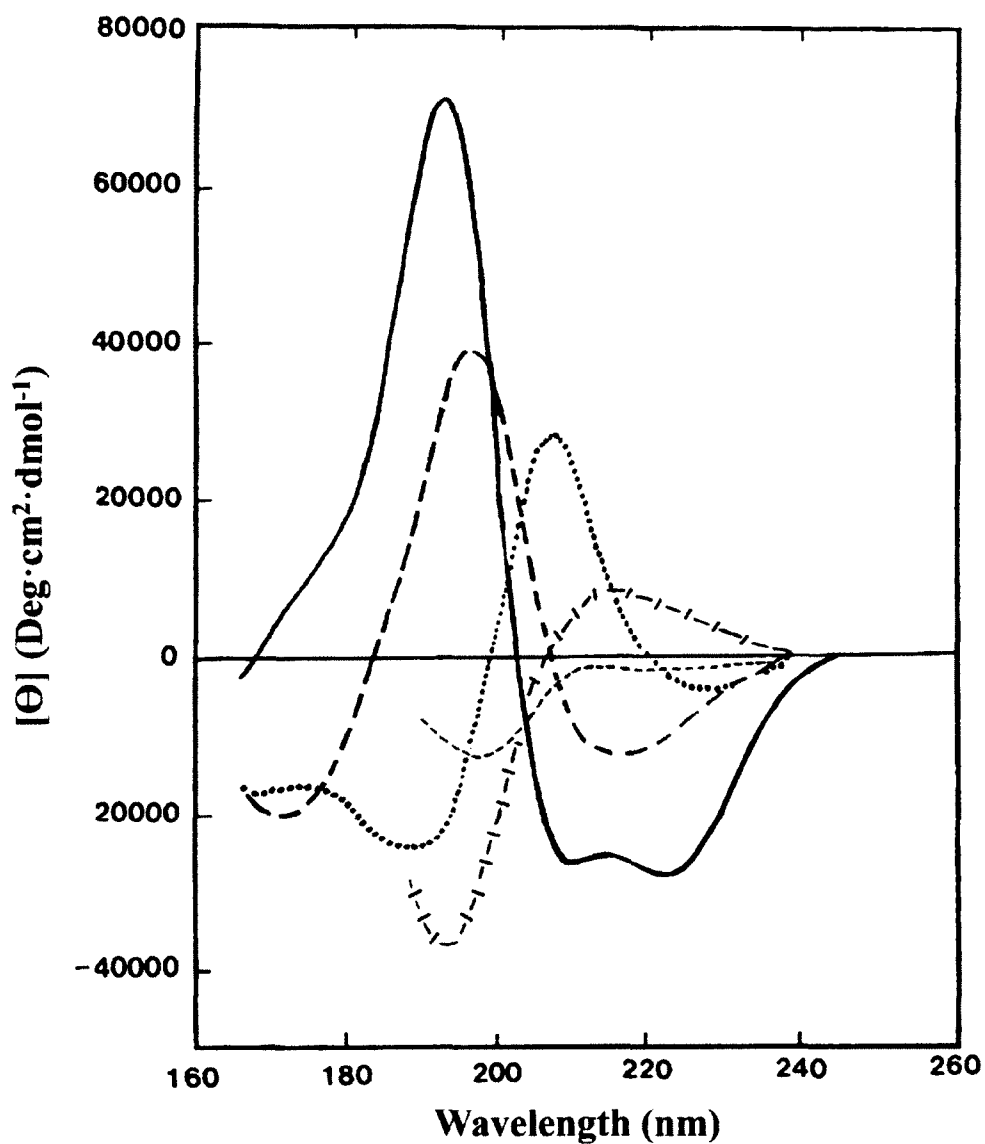


Figure 10. Far-UV CD spectra associated with various types of secondary structure.

Solid line, α -helix; long dashed line, anti-parallel β -sheet; dotted line, type I β -turn; cross dashed line, extended 3_{10} -helix or poly (Pro) II helix; short dashed line, irregular structure. Image reprinted from [139].

In the near-UV region (260-320 nm) the spectra obtained is related to tertiary structure of the protein structure due to the quantity, mobility and environment of the aromatic amino acids. Tryptophan, tyrosine and phenylalanine show peaks in the 290-305 nm, 275-282 nm and 255-270 nm regions respectively (Figure 11) [139, 158]. The structural resolution of the specific aromatic regions is due to the differences in the vibrational excitation states of each aromatic amino acid [139]. Although the near-UV CD spectra are unique to each protein there is information that can be gained though coupling with techniques such as site-directed mutagenesis. Even though the theoretical structural analysis is limited, specific features can be assigned by selectively removing specific aromatic residues and monitoring the effect on the tertiary structure. A few examples of this can be seen in the case of human carbonic anhydrase II, bovine ribonuclease and the molybdate-binding transcription factor ModE from *E.coli* [169-171]. The CD data obtained in the near-UV is quite valuable as a reference fingerprint for the tertiary structure, which is extremely important in identifying molten globule states or when comparing protein variants against the WT structure. Unlike the specific structural elements assigned in far-UV, near-UV analysis provides very little information to the 3D orientation of the structure of a protein.

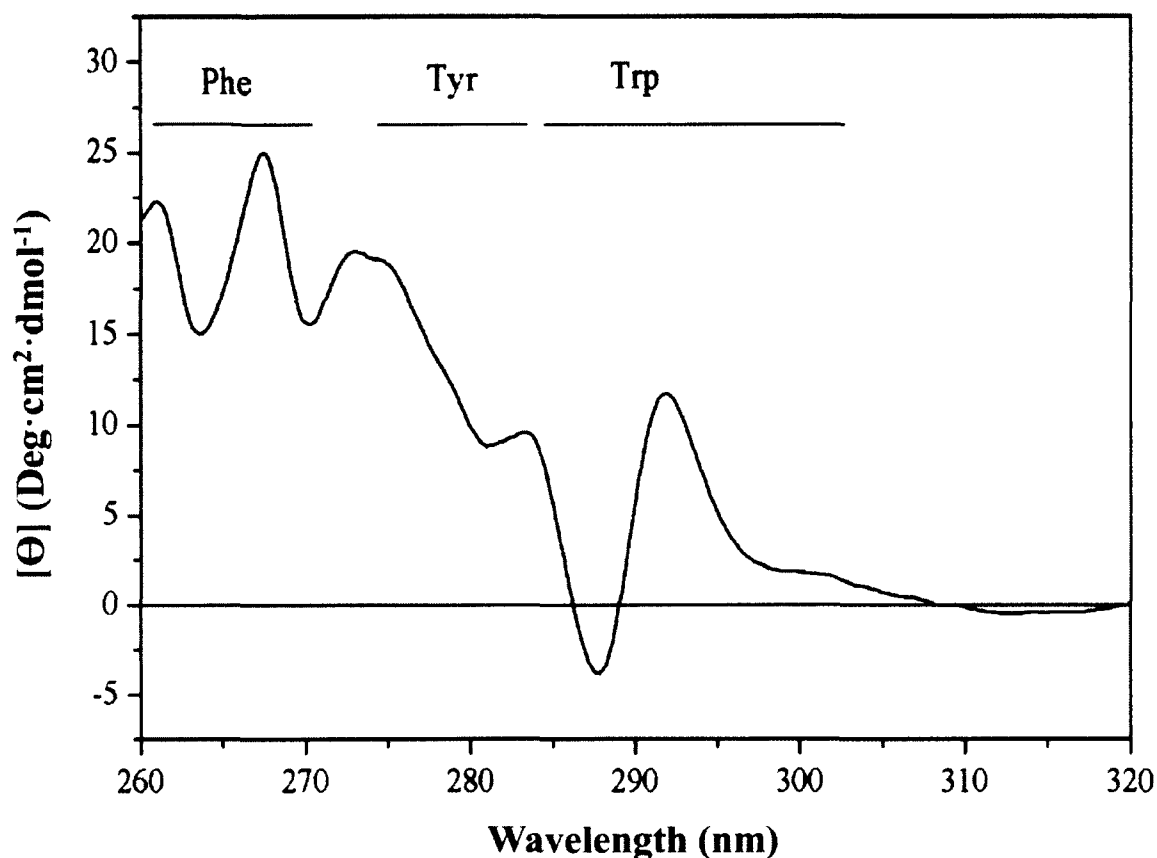


Figure 11. The Near-UV CD spectrum for type II dehydroquinase from *Streptomyces coelicolor*. The wavelength ranges corresponding to signals from Phe, Tyr and Trp side chains are indicated, but it should be emphasized that there can be considerable overlap between the Tyr and Trp signals. Figure adapted from [172].

CD is a versatile method and is extremely important in structural biology, for example as a comparative reference for secondary structure content which is relatable to solving of protein structures using X-ray crystallography and NMR. CD can be used to analyze the stability and changes in protein secondary and tertiary structures as a function of temperature, pH, chemical denaturant and mutation [139]. The advantage to CD spectroscopy is that it is a quick method for gathering structural information on proteins,

on the order of minutes to hours depending on the instrument. In addition, it involves very little protein and does not require extensive post-experiment data processing, allowing a significantly large amount data under various types of conditions to be rapidly obtained. The detailed analysis of structure provided by CD allow for the investigation of conformational changes due to solvent, ionic strength, temperature, pH and the binding of ligands involved in biological function. The rate of change of the secondary and tertiary structure in proteins can be determined in real-time when coupled with a stopped-flow instrument. Because high resolution structural studies like X-ray crystallography are not as amenable to peptides or insoluble aggregates, CD is one of the most predominate methods for studying structural changes associated with amyloid fibril formation [139]. This technique is invaluable in studying the transition from α -helix to β -sheet associated with amyloid fibril formation involved in the misfolding of globular proteins or peptides such as prions [173].

METHODS TO STUDY PROTEIN FOLDING: NUCLEAR MAGNETIC RESONANCE

The NMR spectrometer is composed of a magnet, radiofrequency system, probe and possibly a variable temperature controller (Figure 12) [174]. The magnet stack in NMR of biological molecules is almost universally a helium-cooled persistent superconducting magnet of the highest field strength available. The radiofrequency system is composed of five components; the transmitter, decoupler, power amplifiers, receiver and field-frequency lock system [174]. The transmitter generates broadband excitation pulses at the select nucleus frequency to be observed and can be adjusted to cover all nuclei of interest [174]. The decoupler is a second source of broadband excitation similar to the transmitter. The power amplifiers boost the excitation pulse power to very high wattage levels for non-selective frequency pulses. The receiver detects the free induction decay of nuclear spins and operates over a wide range of frequencies for various nuclei detection [174]. The field-frequency lock system uses an internal reference, typically deuterium, as a constant reference to ensure constant magnetic field. Corrections to the magnetic field are made in response to drift in the reference signal to maintain a uniform magnetic field [174]. Probably the most important component in the NMR spectrometer is the probe. The primary two functions of the probe is to convert the radiofrequency power into oscillating magnetic fields that can be applied to the sample and detection of oscillating magnetic fields generated by the relaxing nuclei [174]. Probes can accommodate different sample sizes although typically 5 mm probes are most common for biological NMR [174].

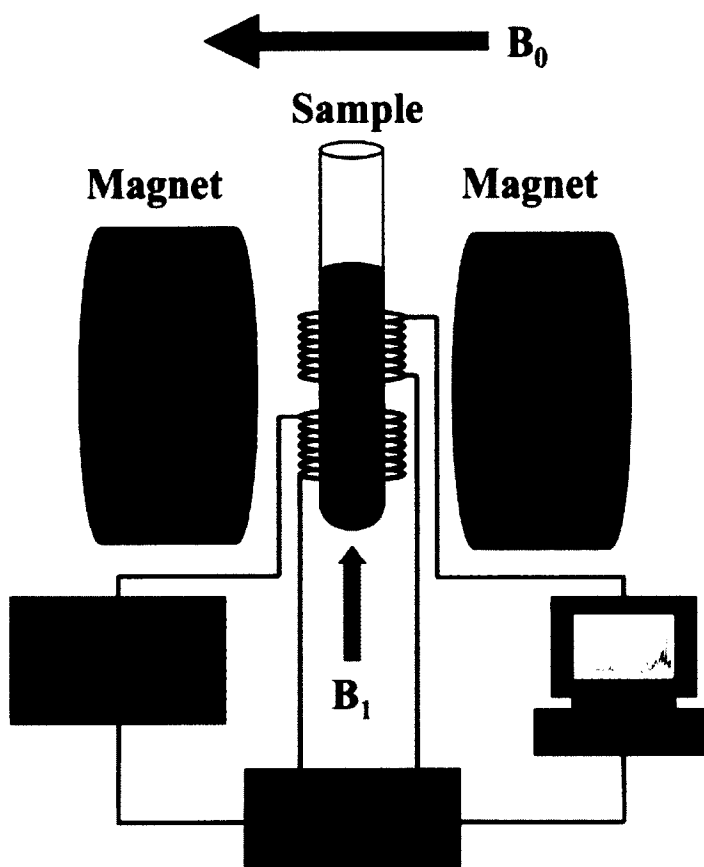


Figure 12. Schematic of an NMR instrument. Shown is the general set-up for conducting NMR with biological samples. Here the macromolecules are placed in an NMR tube which is inserted into the spectrometer. The tube is positioned between two magnets where the atoms of interest are aligned and detected. Image was redrawn from Agilent technologies (http://www.agilent.com/labs/features/2011_101_nmr.html).

NMR is a continually evolving high resolution method for studying complex molecules with atomic-level sensitivity. The NMR spectrum of pancreatic ribonuclease performed on a 40 MHz spectrometer, first reported in 1957, was the first proton NMR analysis of a complex macromolecule. However, the information that could be deduced

from this early analysis was extremely limited. In the next several decades the power of NMR instrumentation and methodology has evolved enormously, providing the ability to investigate conformations and interactions of biological molecules on the atomic scale. The most notable developments include increased sensitivity due to much higher magnetic field spectrometers, development of pulse Fourier transformation methods and multi-dimensional NMR methods. In the most modern applications of NMR, 2D analysis or higher are commonly used to study biological macromolecules. Information of a typical one-dimensional (1D) analysis provides distinct spectral peaks for each atom of interest whereas 2D analysis can provide cross-peak (CRPK) information about connections between atomic resonances. However, the specifics of the information gathered depend on the type of experiment being run. CRPK's can be observed through interactions of bonded atoms up to four atoms apart, interactions of atoms through-space that are within range of each other or through exchange interactions of the same atom in differing environments in the same sample. In example, the dipolar-assisted rotational resonance (DARR) pulse method is used to monitor ^{13}C - ^{13}C interactions through space by a transfer of magnetization (Figure 13) [175, 176]. In this method CRPK's are produced through an initial transfer of magnetization from ^1H to covalently bonded ^{13}C followed by subsequent transfer to ^{13}C carbons that are close in space. Short mixing times allow for intra-residue ^{13}C - ^{13}C CRPK's while longer mixing times allow for inter-residue CRPK's [175, 176].

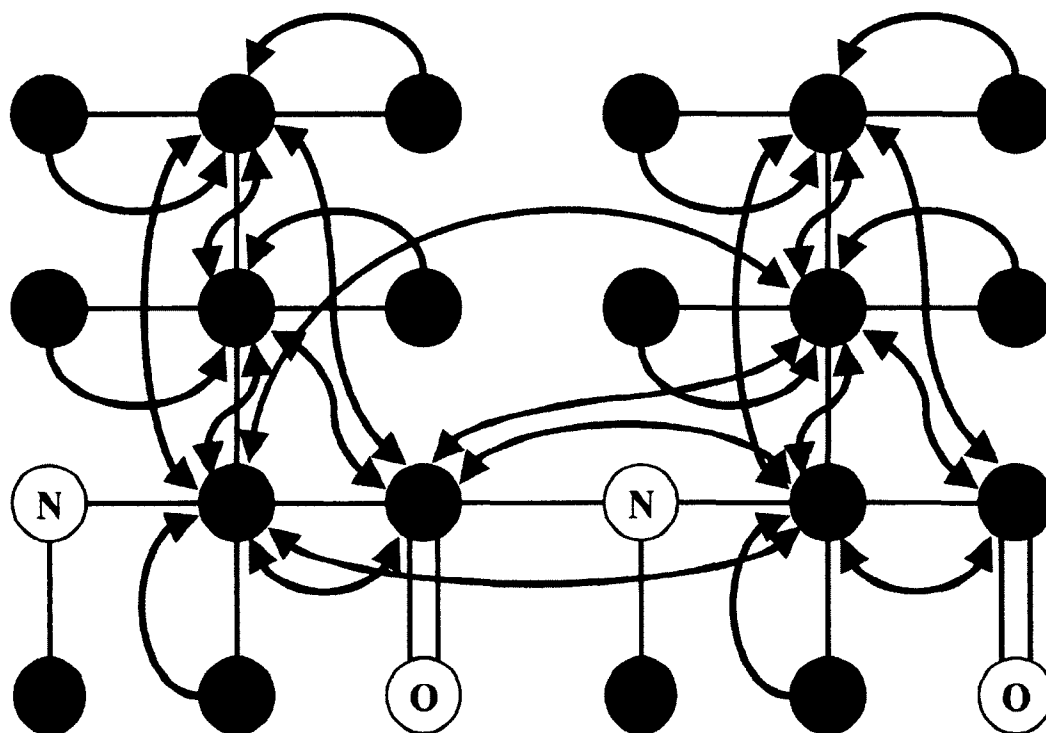


Figure 13. Schematic of the 2D DARR method. This schematic shows the intra-residue (black arrows) and inter-residue (grey arrows) magnetization transfers between ^1H (blue) and ^{13}C (pink). Magnetization transfers between non-covalently bonded atoms result in a CRPK in the 2D spectrum. This figure was redrawn by hand in Microsoft Power Point (V. Higman, University of Bristol).

NMR spectroscopy is possible because atoms of the same element, oriented by a strong magnetic field, produce distinct spectral signals depending on their atomic environment when irradiated with characteristic radiofrequency radiation. The resolution of NMR is very high because the signal is derived from individual atoms and their atomic interactions. The most commonly used isotopes used to study proteins in NMR are ^1H , ^{13}C , ^{15}N and ^{19}F . However, because the natural abundance of the NMR active isotopes of

carbon, nitrogen and fluorine is small in natural biological molecules, typically isotopic labelling can be used to make a biological molecule more NMR sensitive. When studying proteins the first and rate-limiting step of any NMR experiment is the assignment of all the resonances to specific nuclei. However, the relative complexity of the spectrum obtained in the single-dimension and even the second-dimension for proteins is extremely convoluted due to spectral overlap of signals. Moreover, in terms of applications in structural molecular biology, being able to combine high resolution structural data and *in vitro* experimental information makes NMR spectroscopy invaluable in protein folding. In addition, the advance in increased magnetic field strength is continually growing. In the last decade, 700 and 800 MHz magnets are becoming more common and the most advanced spectrometers at 900 MHz and higher are gradually gaining a foothold.

PROTEIN MISFOLDING

In the past decade there has been an increase in the identification of protein misfolding related diseases which include systemic amyloidosis and neurodegenerative diseases [88, 177]. With about 20 different diseases associated with the deposition of amyloid fibrils, there is a need to better understand fibril etiology, pathogenesis, as well as the structure and mechanism of fibrillogenesis. The term amyloid has been defined by the Nomenclature Committee of the International Society of Amyloidosis as extracellular deposits of protein fibrils with the cross- β X-ray diffraction pattern, fibrillar characteristics under electron microscopy and an affinity for Congo red with a resulting apple green birefringence [178]. However, the structural biology and biochemistry community is looking to elucidate the molecular structure, biophysical characteristics and mechanism of amyloid or amyloid-like formation, and have limited the definition to fibrillar polypeptide aggregates with cross- β conformation [178]. Fibrillation results from the rearrangement of a naturally soluble protein or peptide into a stacked β -sheet rich insoluble ordered aggregate filaments (Figure 14), which accumulate into either extracellular deposits in the organs or tissues associated with the specific protein and in some instances leads to amyloidosis [88, 177]. Conversion of proteins into amyloid fibrils is proposed to produce its adverse effect in specific cases by either a loss-of-function or gain-of-function mechanism [88, 177]. It appears the process of fibrillation is not limited to a specific class of proteins, which indicates that the fibrillation process and resulting fibril morphology may be independent of the native secondary and tertiary structure and appears to be an intrinsic property of the primary sequence of amino acids [179, 180]. A fundamental molecular condition for fibrillation is the destabilization of the overall native

structure which allows the primary structure to explore alternative conformations or digestion of the protein resulting in aberrant peptides [88, 177]. It is not clear exactly how fibril aggregates form, but it has been proposed that every protein has the ability to form amyloid fibrils under the appropriate environmental conditions [181]. Interestingly, many proteins not associated with diseases have been shown to form into amyloid fibrils, indicating that fibril formation is an alternative pathway for structural diversity [181-183].

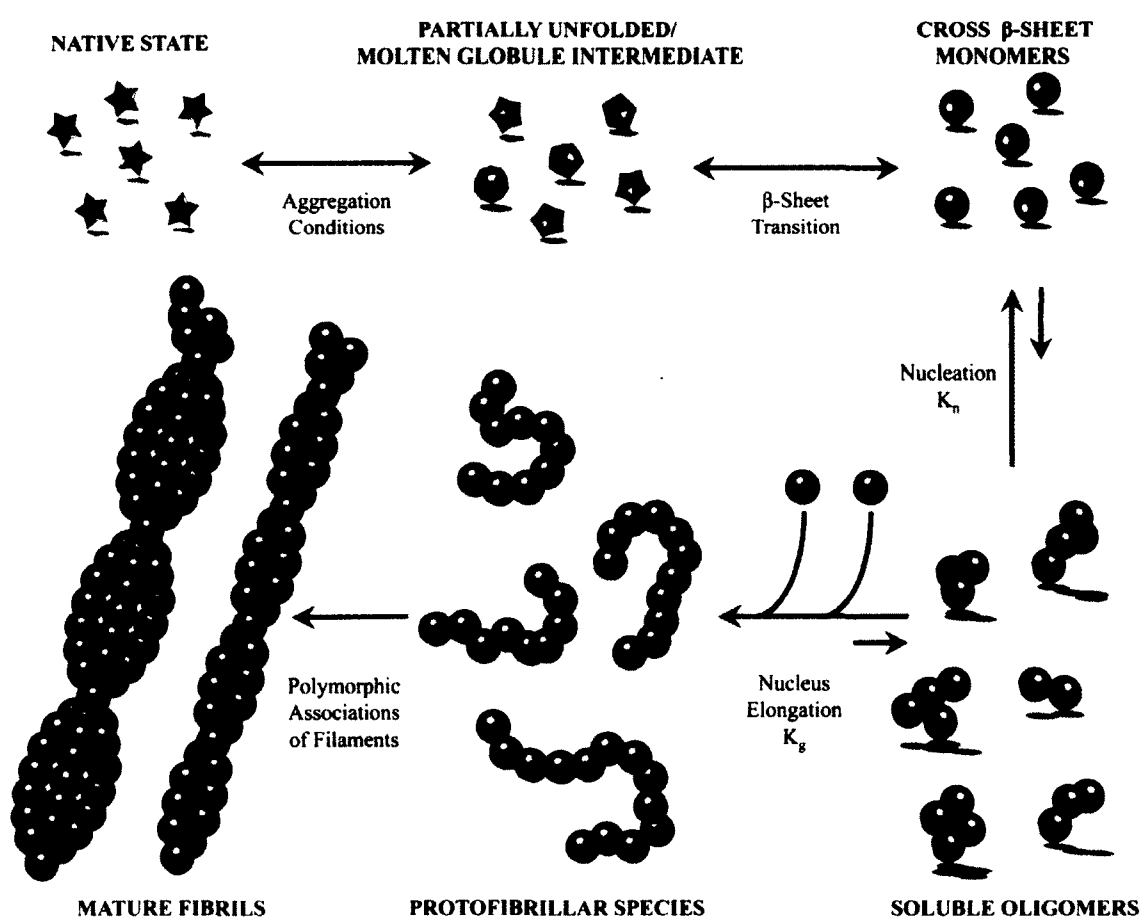


Figure 14. Mechanism of protein amyloidogenesis. Schematic was adapted from concepts in the following papers [184-186] and published in the present reproduced form in [187]. K_n and K_g are equilibrium values of nucleation and growth, respectively.

To better understand the process of amyloid fibril formation we must also better understand the exact key determinants and interactions of protein folding. In normal protein folding the conditions for folding are relatively specific so as to prevent formation of disordered amorphous aggregates and yield a biologically active form [88, 177]. In fibrillogenesis, environmental conditions are similarly specific directing an alternative conformation [88, 177]. In contrast to normal protein folding, the mechanism underlying ordered protein aggregation relies on the destabilization of globular proteins reducing the dynamic stability of the tertiary structure, resulting in the formation of either a molten globule-like state or partially unfolded conformation (Figure 4) [88, 177]. Intrinsically disordered proteins (IDP) do not have structural elements and must first adopt a partial fold before transitioning into amyloid fibrils [177]. The molten globule-like or partially folded conformation allows access to the kinetically alternative folding pathways for conversion into the β -sheet rich amyloid fibril structure [88, 177].

Formation of disease state amyloid deposits *in vivo* typically result from instability in native protein structure or cleavage of peptides, at relevant physiological conditions, due to mutations, environmental changes, chemical inducers or overexpression of proteins which all result in the induction of conformational plasticity [188]. Amyloid fibrils are structurally defined as an unbranched protein polymer with a repeating substructure composed of cross β -sheet structure of indefinite length (Figure 15) [88, 189, 190]. However, it is a multifactorial problem and other factors such as intercellular protein processing also play a role. Amyloid aggregation is distinguished by the formation of the cross- β structure substructure with a 4.7 Å internal strand distance and 10 Å spacing [191, 192]. Recent work revealed two unique types of core interaction

patterns, termed “dry” and “wet” interactions, found within the nucleus of the fibril structure, which indicate that the base unit underlying the amyloid fibril is a pair of stacked β -sheet structures [189]. Differences between these interactions are unique and specific to the amino acid sequence and classified according to amino acid character [189].

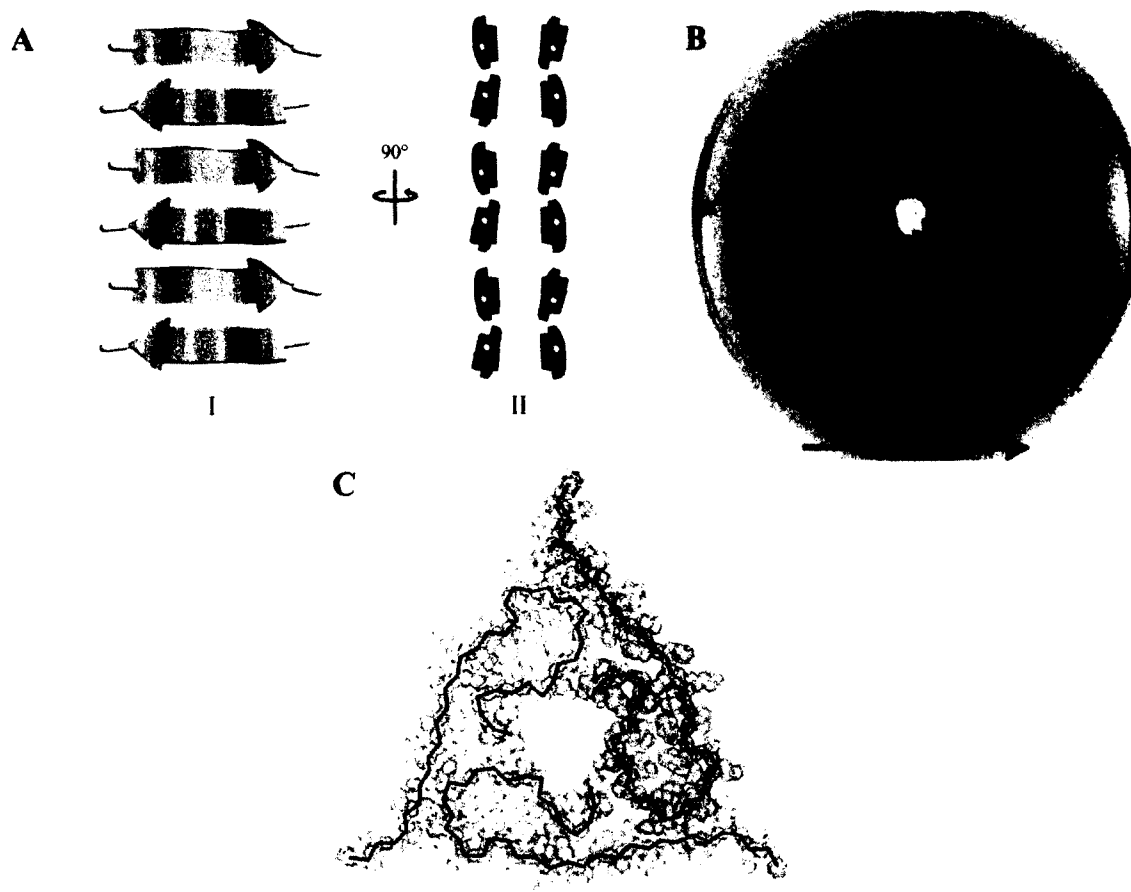


Figure 15. Representative structures of amyloid fibrils and their common X-ray diffraction pattern. (A) Model of fiber forming segments of Aβ with the cross-β diffraction pattern (PDB code: 3OW9) [192], (B) X-ray diffraction pattern of characteristic amyloid cross-β reflections, reproduced from reference [193] and (C) molecular structure of Aβ40 fibrils isolated from Alzheimer's patients (PDB code: 2M4J) [190]. Areas of the backbone colored yellow are indicated as β-strand region. Images A and C were drawn with RasMol (Ver. 2.7.2.1.1).

Fibrillation is not only related to disease states, there are many proteins that adopt a non-cytotoxic amyloid fibril form as a native structure to perform normal biological activities, called “functional amyloids,” which may indicate a fundamental alternative fold in protein evolution (Figure 16) [189, 194]. These functional amyloids are capable of accessing a higher level of protein activity in that they are capable of crossing the boundary of solubility, making them able to perform functions soluble proteins are unable to attain [195]. Functional amyloids can be distinguished from other ordered protein filamental structures by the unique formation of cross- β sheet structure, whereas structures found in the assembly of actin filaments, microtubules, fibroins, some silk, collagens and keratins do not form the cross- β motif [196-199]. Transition to or from the functional amyloid form can regulate biological activity in which there is interconversion between active and in-active amyloid form. [200-202]. In the case of HET-s prion protein the transition is accompanied by a gain of protein function, whereas the yeast prion protein is accompanied by a loss of function [200-202]. Functional amyloids have been shown to perform a variety of functions such as ligand binding, extracellular adhesion, biofilm formation, sorting, storage and release of hormones (Table 1). Interestingly, of 42 peptide and protein hormones in the secretory pathway, 31 form amyloid fibrils *in vitro* under either granule-relevant pH alone or in the presence of heparin, a common granule glycosaminoglycan (GAG) and are biologically necessary for the development and survival of many organisms [195]. It is curious to find that fibrillation of a hormone would be a functionally relevant form as the amyloid fibril is considered to be a very stable low energy form [191]. Meaning that the functional monomeric hormones would be locked in the fibril substructure remaining non-functional

and disassembly of the fibril structure must be a condition required for biological activity. This requirement is met by amyloid hormones upon dilution into the extracellular matrix releasing monomeric biologically active hormones [195]. The underlying mechanism of amyloid disassembly is still not completely understood. However, the ability of proteins or peptides in the secretory pathway to form densely packed granules containing insoluble amyloids followed by secretion and disassembly into the extracellular space is significantly important in the evolution of the amyloid form. This transition appears to be regulated and controlled environmentally by changes in pH as well as presence of a cofactor GAG [195]. This reversible conversion of hormone proteins into amyloid fibrils as a packaging, storage and release mechanism in the secretory pathway appears to be a uniquely evolved system that does not discriminate between proteins of differing secondary or tertiary structures. This function is unique and may hold key information in conditions and requirements required for the disassembly of highly stable amyloid fibrils.

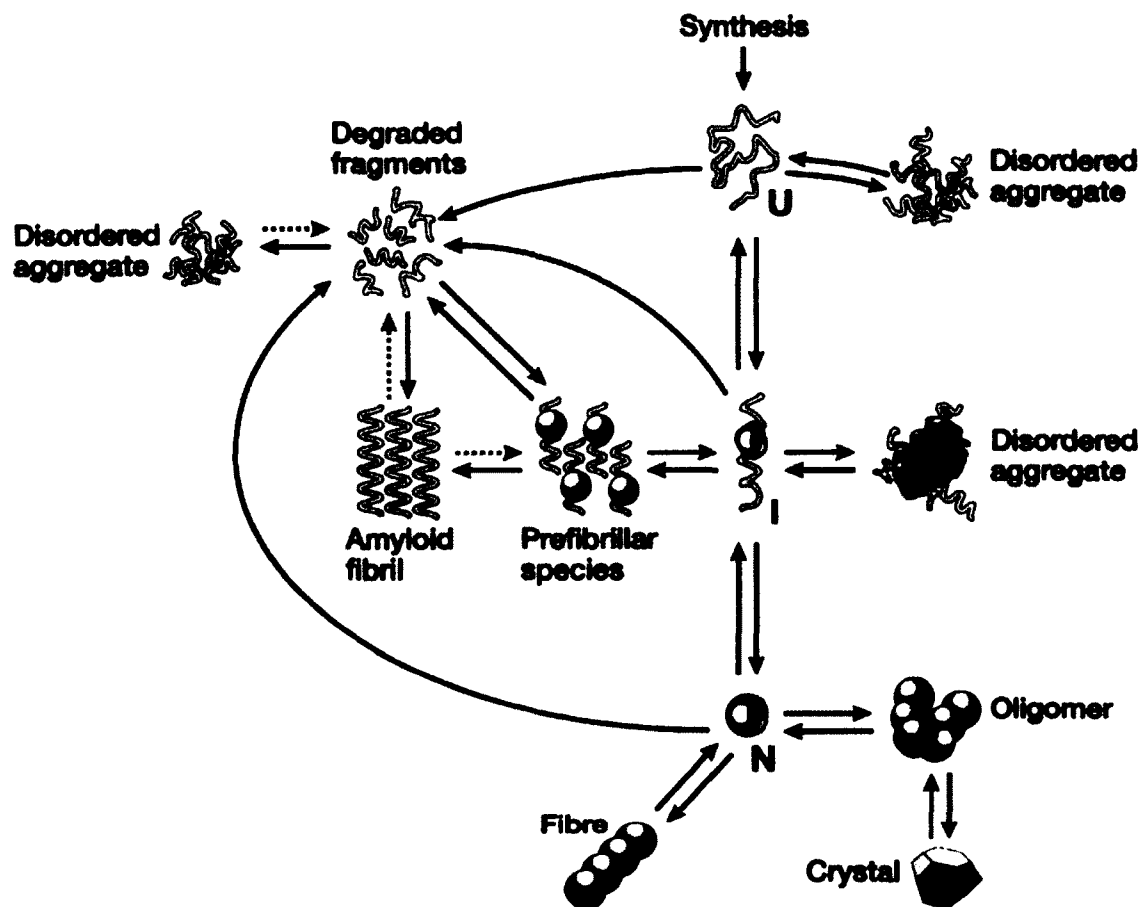


Figure 16. Schematic that represents the possible polypeptide fates following translation. Each fate is tightly regulated by various control mechanisms and the formation of each state is dependent on the thermodynamic and kinetic accessibility requirements of formation. Reproduced from reference [203].

Table 1. Examples of small α -helical polypeptide hormones known to be functional amyloids. Adapted from [195].

Protein/Peptide	Organisms	# of Residues	Tissue Type
Corticotropin Releasing Factor	Rat/Human	41	Hypothalamus
Urocortin II	Human	38	Hypothalamus
Urocortin III	Human	38	Hypothalamus
Glucagon	Human	29	Pancreas
Glucagon-like Peptide 1	Human	37	Pancreas
Glucagon-like Peptide 2	Human	33	Pancreas
Gastric Inhibitory Polypeptide	Human	42	Gastrointestinal tract
Exendin-4	Lizard	39	Salivary Gland
Neuromedin K	Porcine	10	Spinal cord
Neuropeptide Y	Porcine	36	Hypothalamus

The sequence of amino acids dictates the fold of a protein and a part of solving the protein folding problem surrounds the question; can we predict the fold of a protein from its primary structure? There is a very large area of research dedicated to understanding the sequence determinants of amyloid fibril formation and developing prediction methods which is the flip-side of the native protein folding problem. It is thought that residues important in fibrillation are different from those found in normal protein folding even though there appears to be cross over in residues that play critical roles in hydrogen bonding and forming hydrophobic surfaces [204]. There is a clear indication that specific sequences of amino acids tend to have a higher propensity to aggregate even though most of the polypeptide chain can be a part of the fibril substructure [188, 205]. In comparison to the nucleation or hydrophobic collapse models of protein folding where a specific set of amino acids form a critical folding nucleus or

hydrophobic core respectively, short sequences of amino acids have been shown to induce aggregation of larger proteins that were otherwise normal, indicating the possibility of a similar set of critical amino acids that govern fibrillogenesis [32, 206, 207]. It was efficiently described using mutational analysis of acylphosphatase, that the chemical characteristics, concentration and environmental conditions of the polypeptide chain dictate the rate and propensity for fibril formation [208, 209].

Today, prediction programs such as TANGO and WALTZ can be used to determine the propensity of the entire polypeptide chain or possible patches of residues within a given sequence of amino acids to be susceptible to amyloid fibril formation respectively [210, 211]. Aggregation prediction further support the mechanistic need of globular proteins to form partially unfolded intermediates before fibrillation can occur because aggregation prone regions are not unique to a single type of secondary structure or location within the protein [191]. However, it should be noted that a protein that is misfolding can transition down either aggregation or fibrillation pathways exclusively, although in some instances aggregation can precede fibril formation. Despite the success of these programs they are still incomplete and further development of these algorithms to better hone their predictive capabilities is required for aggregation prediction *in vivo*. Analysis of full genomic proteins using the aforementioned prediction methodology revealed an important separation in protein evolution [212, 213]. Sequences of amino acids encoding functional proteins are significantly less amyloidogenic when compared to random sequences of amino acids, indicating that some external stimuli have adapted proteins to avoid the amyloid form [212, 213]. It is normal to assume that IDP's would be more aggregation prone but in fact they are less susceptible than globular proteins to

ordered aggregation [214]. It appears that nature has evolved internal prevention measures such as placing β -sheet resistant or charged residues near sequences of amino acids that are prone to β -sheet aggregation, using proline residues in helical membrane proteins to protect from improper folding, β -bulges to protect edge strands in native β -sheet proteins or specifically conserved glycine, all resulting in an overall reduction of amyloid propensity [213, 215-217]. Experimental evidence of these evolutionary designs can also be found in globular proteins such as β 2-microglobulin (β 2M), transthyretin (TTR), superoxide dismutase (SOD), acylphosphatase and fibronectin [217-221].

However, in β 2M prolonged renal dialysis as well as mutations or truncations in TTR and mutations SOD can disrupt these designs and induce fibrillation [222-226]. It is believed that the protective nature of the designs mentioned above functions in increasing or introducing new high energy barriers that prevent the formation of intermediate unfolded or molten-globule like states required in the conversion to β -sheet rich aggregates [218]. It is imaginable that repetitive sequences of amino acids that alternate in polar and non-polar character would inevitably result in rapid amyloid aggregation, but nature has developed proteins to avoid this sequence, further reducing aggregation propensity while maintaining the ability to fold into a functional protein [227]. It is truly amazing how proteins have evolved to avoid amyloid fibrillation while still maintaining key features that allow for rapid folding to the functional native state.

In this section we will discuss the fascinating fibrillation process with all proteins that share the all α -helical secondary structure. This will be a unique opportunity to investigate similarities and differences in specific conditions required for the formation of amyloid fibrils within a specific class of proteins. Understanding how an all α -helical

protein transitions into β -sheet amyloid fibrils is also of significant importance because it is the most extreme case of significant structural modifications required for fibrillation through an α -to- β conversion. Depositions of several α -helical proteins are associated with known disease states (Table 2). Four of these all α -helical proteins are apolipoprotein A-I, insulin, lung surfactant protein C (LSP-C) and prolactin (Figure 17) [88, 180]. Deposition of these fibrils can be deleterious to the life and function of an organism so here we will discuss the α -helical proteins whose amyloid deposition is associated with a disease state (Table 2). Formation of disease state amyloid deposits *in vivo* typically result from instability in native protein structure, at relevant physiological conditions, due to mutations, environmental changes or overexpression of proteins which all result in conformational plasticity. It has been established that the *in vitro* formation of amyloid fibrils requires the destabilization of the native state of the protein by extreme conditions, mutations or chemical inducers. Mature fibrils found in clinical plaques are typically found associated with many different species such as GAG's, metals and lipoproteins to name a few (61, 62). However, gross-morphologically similar amyloid fibrils can be formed from *in vitro* isolation of native protein or recombinant native proteins in the absence of binding species (i.e. GAG's, metals or lipoproteins) using destabilizing conditions to accelerate formation (1). Five all α -helical proteins; the androgen receptor protein, apolipoprotein A1, insulin, LSP-C and prolactin protein have links to disease states [88, 180]. Thus, understanding the transition of all α -helical proteins into amyloid fibrils has been critical to better understanding the key determinants of amyloid fibril formation.

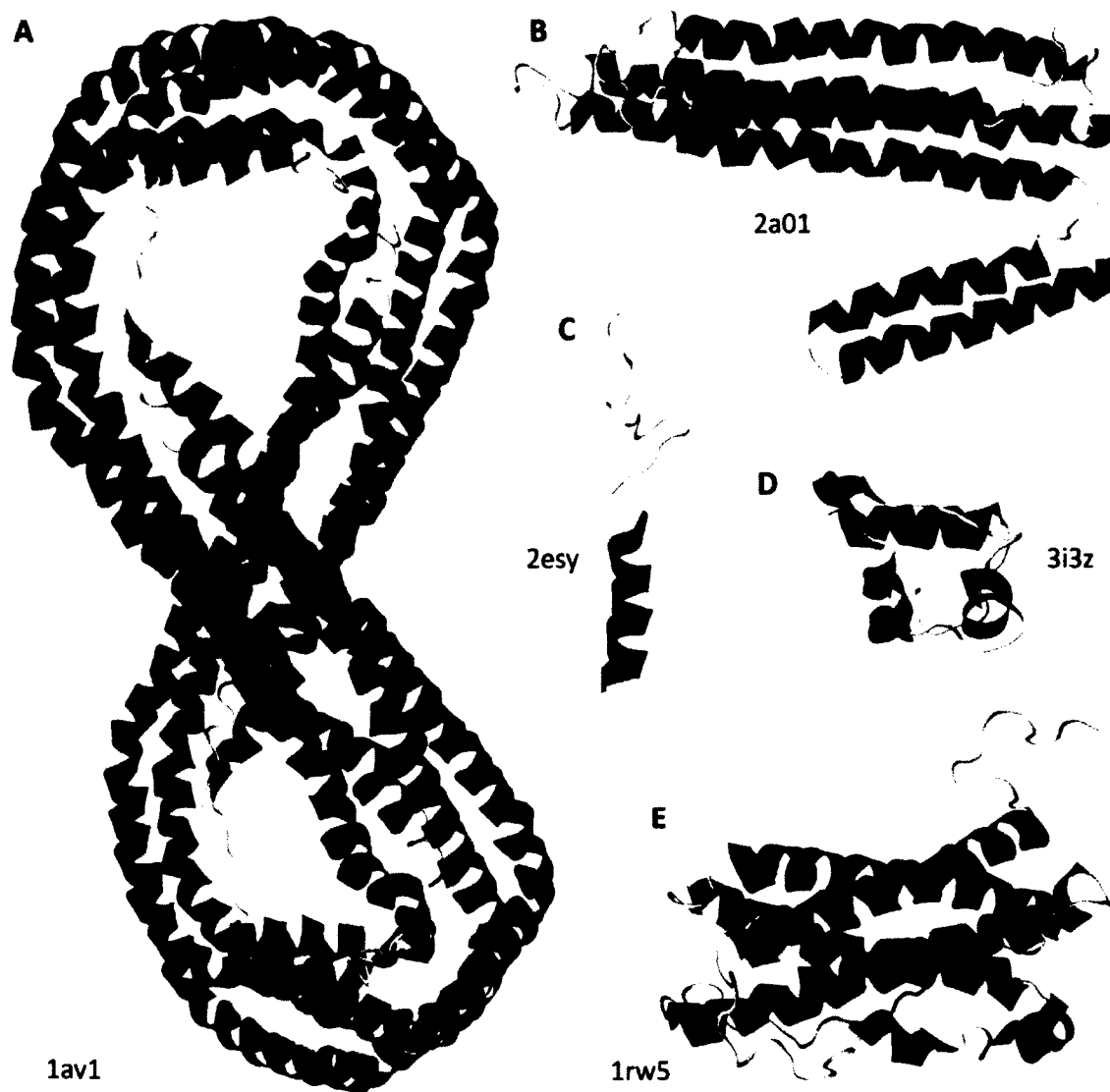


Figure 17. Molecular structure of α -helical proteins related to amyloidogenic diseases. The α -helical proteins shown are a truncated homotetramer form of apolipoprotein A-I (**A**), monomeric apolipoprotein A-I (**B**), lung surfactant protein C (**C**), a monomeric insulin-like molecule (**D**) and human prolactin (**E**). The PDB codes are shown next to each structure. Images were made in Jmol (Ver. 13.0.4.).

Table 2. List of all α -helical proteins known to form amyloid fibrils clinically, *in vivo* or *in vitro*. Adapted from [180].

Disease Related all α -helical Proteins

Protein/Peptide	Disease	Fibrillation Conditions	Time (Days)
Apolipoprotein A-I [228]	Hereditary systemic amyloidosis (i.e. atherosclerosis)	100 mM Tris-HCl buffer (pH 7.4) at 37 °C	10
Insulin [229]	Injection-site amyloidosis	HCl (pH 1.6) at 60 °C and 70 °C	7
Lung surfactant protein C [230]	Pulmonary alveolar proteinosis	CHCl ₃ /CH ₃ OH/0.1 M HCl at 37 °C	14
Prolactin [231]	Age-related pituitary amyloidosis	5% D-Mannitol (pH 5.5), 0.4 mM chondroitin sulfate A, at 37 °C, 50 rpm agitation	30

Non-Disease Related all α -helical Proteins

Protein/Peptide	Fibrillation Conditions	Time (Days)
Apoptotic protease activating factor-1 caspase activating and recruitment domain [232]	Glycine-HCl buffer (pH 2.1) at 60 °C	4
Apomyoglobin [182]	Na ₂ B ₄ O ₇ buffer (pH 9.0) at 65 °C	25
Bovine cytochrome <i>c</i> [233]	Tris-HCl buffer (pH 9.0) at 75 °C	0.5
Bovine serum albumin [234, 235]	Tris-HCl buffer (pH 7.4), 75 °C; Glycine-HCl buffer (pH 3.0), 50 mM NaCl, 65 °C	12.5
Cytochrome <i>c</i> ₅₅₂ (Cys11Ala/Cys14Ala) [236]	Physiological Conditions (pH 7.3)	56
Fas-associated death domain [183]	20 mM Glycine-HCl buffer (pH 2.1), 150 mM NaCl, at 50 °C, 180 rpm agitation	34
α -helical hormones [195]	5% D-Mannitol (pH 5.5), with or without 0.4 mM heparin, at 37 °C, 50 rpm agitation (0.01 % azide)	30

Since the discovery that proteins not associated with disease or formed *in vivo* can undergo the fibrillation process there have been several α -helical proteins that have been induced to form amyloid-like fibrils *in vitro* (Table 2). It is also of interest to note that there are a number of mainly α -helical proteins that have also been shown to convert into an amyloid-like form such as lysozyme [237]. Additionally, there are a select number of IDP and peptides, such as A β , α -synuclein and calcitonin that form all α -helical structures upon exposure to additives (for example, PG vesicles, metals, fluoroalcohols or SDS micelles). We intend to specifically review the process for α -helical proteins to form aggregated species and the transition to amyloid fibrils and discuss any patterns that may have emerged.

Alpha-helical proteins associated with disease states

Apolipoprotein A-I

Blood contains many transport proteins that move small molecules throughout the body. The deposition of either WT apolipoprotein A-I (Figure 17A) or N-terminal fragments (Figure 17B) has been linked to several systemic forms of amyloidosis such as age-related pulmonary artery amyloid found commonly in dogs or atherosclerotic plaques found in humans [238-241]. Apolipoprotein A-I belongs to a group of plasma exchangeable apolipoproteins that functions in cholesterol transport in the circulatory system in either high density lipoprotein bound or the less abundant lipid-poor/lipid-free forms [242, 243]. Approximately 19 mutations of apolipoprotein A-I have been associated with different types of hereditary and non-hereditary forms of amyloidosis [240, 241, 244-249]. These fibrils are found in senile plaques and have been linked to

interactions with amyloid- $\beta_{(1-42)}$ indicating possible involvement in Alzheimer's pathology [228]. Recent research has identified a short N-terminal sequence of residues 46-59 of apolipoprotein A-I that may be responsible for fibrillogenesis of the full length protein [250]. It is worth mentioning that fibrillation of apolipoprotein A-I under physiological conditions may occur because of intrinsic conformational instability due to natural methionine oxidation and/or the ability to form higher order oligomers which confer structural plasticity [251-253].

Insulin

Insulin is a small 51 residue covalently linked α -helical heterodimer (Figure 17D) which functions in the regulation and metabolism of glucose [254, 255]. Soluble insulin forms higher order multimeric species under relevant physiological conditions to facilitate storage [256]. Association of insulin monomers is regulated by a variety of environmental conditions (for example, pH, concentration, ionic strength and presence of metals) [256, 257]. Dysfunction in insulin production resulting in improper regulation of glucose is considered one of the hallmarks of diabetes. Improper regulation of insulin in diabetes patients requires frequent injections of insulin as the predominant treatment. It seems that the deposition of insulin fibrils is possible in humans, either obtained naturally by aging, diabetic dysfunction or by surface deposition at insulin injection sites [256, 258-260]. The mechanism of insulin fibril formation *in vivo* is believed to require the reduction of higher ordered insulin species to the monomeric form before proceeding as expected with a destabilization of the native structure in favor of a partially unfolded intermediate [180, 256]. Fibrillation of insulin *in vitro* can occur under variety of conditions but typically require high temperatures, organic solvents, acidic pH and

hydrophobic surfaces [261, 262]. The kinetics of insulin amyloid-like aggregation can be modulated by accelerants such as urea or inhibitors such as trimethylamine N-oxide [261]. Modulation of the hydrophobic core of insulin seems to be a critical key in inducing amyloid fibrillation [256]. Also differences in amino acid composition between differing species, such as differences between bovine and human insulin sequences, affect the aggregation kinetics of insulin [263].

Lung surfactant protein C

LSP-C (Figure 17C) plays a role in the very rare pulmonary alveolar proteinosis in which accumulation of surfactant protein amyloid-like fibrils impair respiration [230]. Etiology of this rare disease has been associated with improper clearance of aberrant surfactant proteins by macrophages [230]. LSP-C is one of the most hydrophobic peptides known to exist naturally and functions in concert with a complex mixture of proteins and phospholipids in the lungs to reduce surface tensions at the alveoli interface [264]. This small protein is believed to be post-translationally cleaved from a larger precursor protein upon insertion into a lipid-bilayer [264, 265]. Hydrogen/deuterium exchange experiments show that the transition of α to β occurs through short lived and almost completely unfolded intermediate [264]. Despite destabilizing conditions, this transition is relatively slow due a significant energy barrier associated with the nature of LSP-C's rigid α -helical structure [265]. It is interesting to think about how nature has dealt with the polyvaline sequence in LSP-C which is prone to β -sheet formation [264, 266]. Even though the structure of LSP-C is very stable it is of importance to note that the precursor form may be more stable when not embedded in the lipid-bilayer and may indicate an evolutionary adaptation to prevent the aggregation of the active protein until

cleavage of the precursor protein occurs [264, 265]. Also palmitoylation of two N-terminal cysteine's may also indicate further evolutionary selection for native structure stability because depalmitoylation of LSP-C reduces the stability of the α -helical structure, decreases the energy barrier for unfolding and increases the fibrillation rate of LSP-C [230, 264, 265]. It appears nature has evolved this highly conserved small peptide in many facets most likely due to its high importance in proper lung function.

Prolactin

The all α -helical 23 kDa hormone prolactin (Figure 17E) contains 199 amino acids in its mature form and is predominately, but not limited to secretion by the pituitary gland [267]. Prolactin functions in a significant number of biological processes (>300) making it one of the most diverse functioning peptide hormones [267]. Prolactin post-translational modifications such as polymerization, phosphorylation and glycosylation provide functional plasticity [267]. Prolactin seems to be associated with group of functional amyloid hormones that use transient aggregation mechanisms to store high hormone concentrations in secretory vesicles [195, 231]. Aggregation of prolactin in secretory granules is believed to be a more efficient method for hormone release [195, 231]. In an investigation of spontaneous aggregation of functional amyloid hormones, prolactin spontaneously formed amyloid fibrils in the presence of a unique prolactin GAG (Table 1) [195]. Amyloid structure was verified by specific binding of Thioflavin T (ThT) and Congo red (CR), transmission electron microscopy (TEM) imaging and X-ray diffraction which will be discussed later in this chapter [195]. There are a number of other small α -helical hormones (Figure 18) that have been shown to form amyloid fibrils under similar conditions as prolactin however, with different GAG's (Table 2) [195].

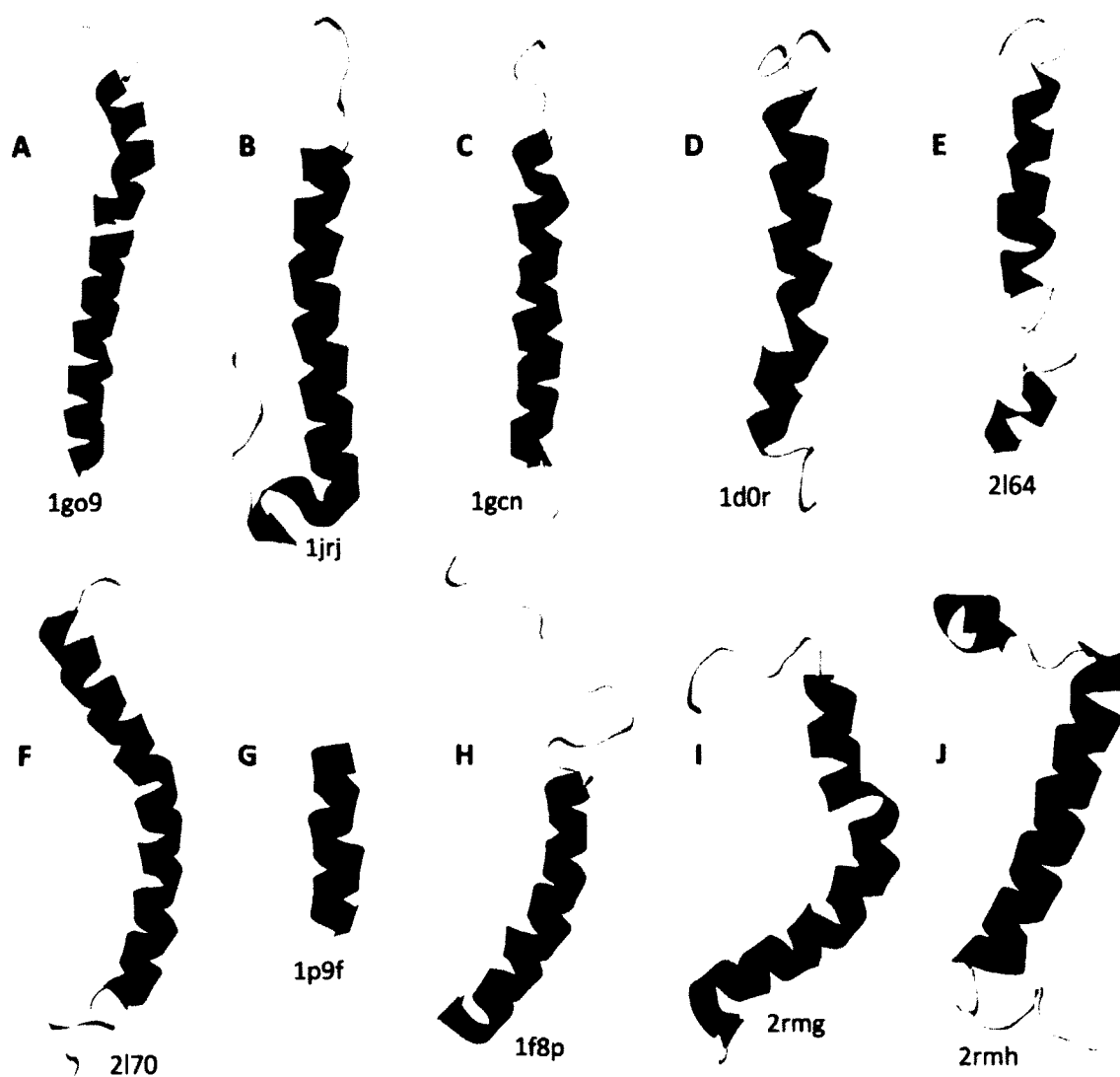


Figure 18. Molecular structure of small hormone α -helical proteins. Hormones shown are corticotropin-releasing factor (L-Phe12 to D-Phe/Leu15 variant) (A), exendin-4 (B), glucagon (C), glucagon-like polypeptide 1 & 2 (D & E), glucose-dependent insulinotropic polypeptide (F), neuromedin K (G), neuropeptide Y (H), urocortin II and III (I & J). The PDB codes are shown next to each structure. Images were made in Jmol (Ver. 13.0.4.).

Alpha-helical proteins not associated with disease

All α -helical proteins that function in apoptosis

Apoptosis can be activated by either intrinsic or extrinsic stimuli both resulting in programmed cell death via release of protein specific proteases [268]. In the apoptotic cell signaling pathway, there are two all α -helical Greek-key proteins which are members of the death domain superfamily, apoptotic protease activating factor 1 (Apaf-1) with an N-terminal caspase recruitment domain (CARD) (Figure 19A) and the death domain of the Fas-associated death domain (Fadd-DD) (Figure 19B), that have been induced to form fibrils *in vitro* [183, 232].

The multi-domain Apaf-1 CARD protein functions as an activator of the caspase cascade in response to intrinsic cellular stimuli such as endoplasmic reticulum stress or DNA damage [269, 270]. This protein has been shown to fibrillate under a specific set of extreme conditions (Table 2). The transition had a lag phase of ~9 h followed by a typical exponential elongation phase and maturation after ~80-100 h [232]. This process of amyloid fibril formation for Apaf-1 CARD appears to be pH dependent and proceed by possible rearrangement of a destabilized molten globule like structure [232]. At pH 4 there are structural changes that indicate the presence of a molten globule-like state of the protein that may be a possible intermediate conformation that facilitates structural rearrangement and formation of protofibrillar oligomers under lower acid conditions [232]. Interestingly, the addition of NaCl for ionic strength appears to stabilize the molten globule state by reducing surface electrostatic repulsions, quite possibly reducing the ability to alter conformation. At lower pH (pH 2.1) in the absence of NaCl there is an indication that the molten globule-like protein rearranges in a manner that reduces

hydrophobic surfaces and forms precursor aggregates, which are capable of forming protofibrillar structures followed by mature amyloid-like fibrils [232].

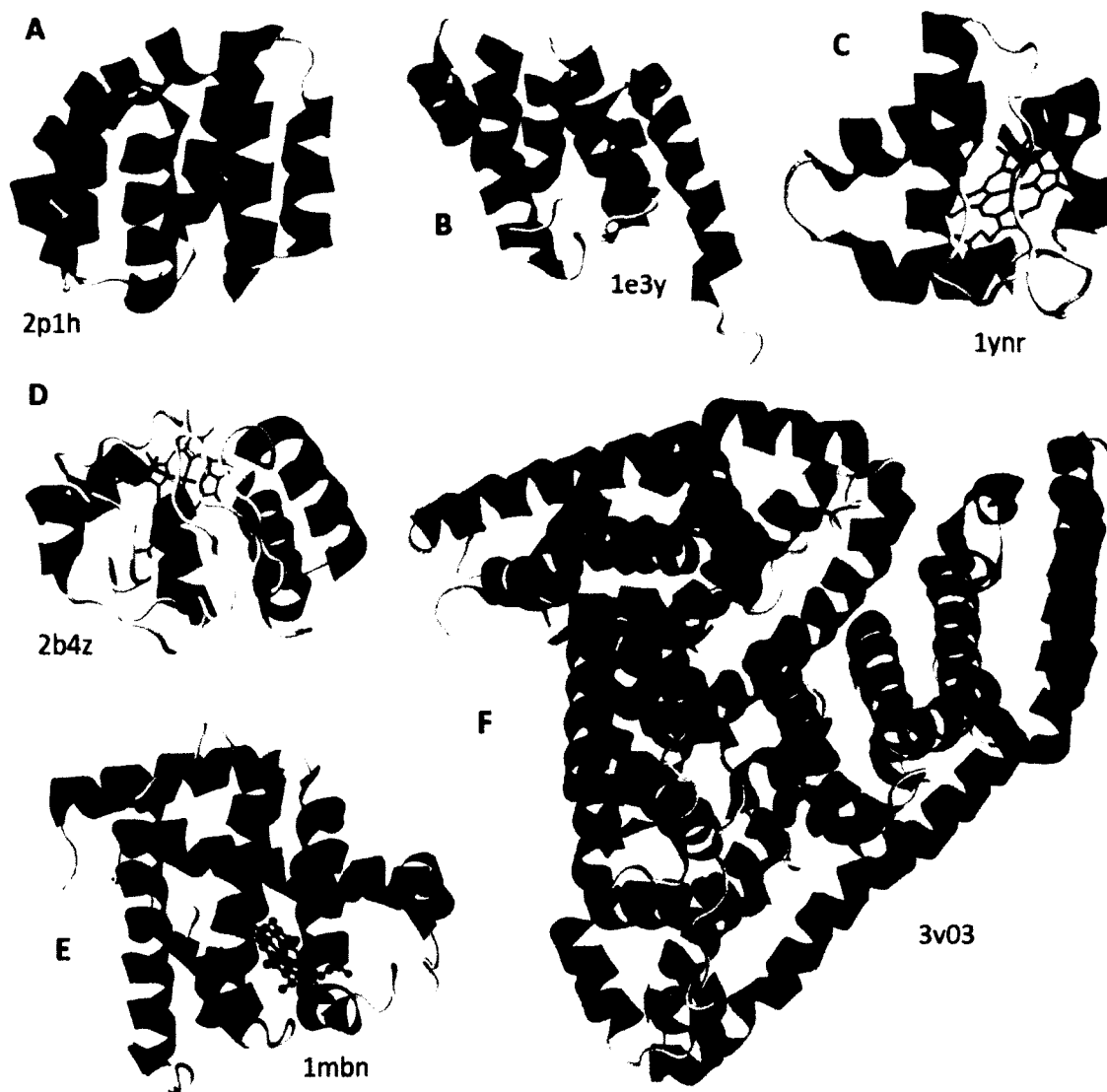


Figure 19. Molecular structure of α -helical proteins not related to disease. The α -helical proteins shown are (A) apoptotic activating factor-1 caspase activating and recruitment domain, Fas-associated death domain (B), cytochrome c_{552} (C), bovine cytochrome c (D), myoglobin (E) and bovine serum albumin (F). The PDB codes are shown next to each structure. Images were made in Jmol (Ver. 13.0.4.)

The Fas-associated death domain proteins are adaptor proteins that facilitate the formation of the death-inducing signaling complex in response to external stimuli [270]. They appear to play a prominent role in the apoptotic pathway initiated by A β plaques in Alzheimer's disease [271, 272]. Interestingly, the Fadd-DD protein also has the ability to form into amyloid-like fibrils *in vitro* [183]. The condition for amyloid-like fibrillation of the Fadd-DD protein is similar to that of Apaf-1 CARD requiring similar buffer and pH. However the conditions deviate significantly in ionic strength, temperature and mechanical agitation. The fibrillation of Fadd-DD seems to proceed through a destabilization of the native structure resulting in a partially unfolded state that still has α -helical composition [183]. Unlike Apaf-1 CARD, the need for ionic strength seems to be critical in the formation of Fadd-DD amyloid-like fibrils by possibly stabilizing electrostatic interactions that might cause unordered aggregation [183]. Since amyloid fibril formation is considered an alternative low energy conformation unique to the polypeptide chain, how have proteins avoided this evolutionary conformational level? It appears that these α -helical proteins have evolved in a manner where amyloid formation is avoided by requiring extreme conditions for fibrillogenesis [270].

Prosthetic all α -helical proteins

Respiration via the mitochondria is present in all eukaryotic organisms. Two small all α -helical *c*-type cytochrome proteins both functioning as electron transport molecules have been shown to form amyloid fibrils under unique methodologies. Both cytochrome *c*₅₅₂ from *Hydrogenobacter thermophilus* (Figure 19C) and bovine heart cytochrome *c* (Figure 19D) covalently bind a haem group in their native structure and contain 80 amino acid and 104 amino acids, respectively. In the case of cytochrome *c*₅₅₂

Cys11Ala/Cys14Ala variant there is a clear structural destabilization of the native structure upon loss of the haem group [236]. This destabilization allows this small all α -helical protein to form into amyloid-like fibrils under physiological conditions (Table 2) [236]. The presence of thioether linkages to haem may indicate why cytochrome *c*₅₅₂ does not form amyloid fibrils *in vivo* and an evolutionary step in avoiding amyloid fibrillation. However, another approach using bovine cytochrome *c* indicate that damage, but not removal, of the haem group also results in the formation of amyloid aggregates as the protein adopts a predominant random coil to allow for conformational rearrangement [233].

Another prosthetic protein not related to a disease that has been shown to form amyloid-like fibrils *in vitro* is myoglobin (Figure 19E). The most commonly described function of myoglobin is the storage of dioxygen in muscles, however it has also been described in nitric oxide scavenging and as a hypoxic nitrite reductase [273-275]. After removal of its haem group, apomyoglobin was induced to form amyloid fibrils under basic conditions found in Table 1 [276]. Interestingly, apomyoglobin maintains its helical content after removal of the haem group under mild conditions (22 °C), whereas at 65 °C apomyoglobin dramatically destabilized resulting in fibrillation [276]. However, investigation into this process revealed that at higher temperatures (i.e. 90 °C) fibrillation is significantly disrupted whereas at lower temperatures (i.e. 50 °C) protein folding appears to suppress the formation of amyloid fibril structures [182]. Mutations of WT myoglobin such as Trp7Phe/Trp14Phe destroys the ability of myoglobin to bind its haem group and Val10 significantly reduces its stability allowing this α -helical protein to form amyloid fibrils at physiological pH [277-279]. It appears that the evolution of protein

function with prosthetic groups such as haem seems to also function as a stabilizing factor that prevents these types of protein from forming amyloid fibrils *in vivo*.

Serum albumins in plasma

Serum albumins are large multi-domain all α -helical plasma proteins that function in the transport of small molecules (i.e. metals, hormones, fatty acids and drugs) to specific tissues, regulate osmotic pressure, maintenance of blood pH and serve as predominant plasma antioxidants [280, 281]. Bovine serum albumin's (BSA) triangular structure (Figure 19F) is composed of three homologous domains with two subdomains each, stabilized by 17 disulfide bridges that provide rigidity but still allow for functional flexibility [234, 280, 281]. BSA was optimally formed into ordered amyloid-like fibrils with a cross β -sheet structure under the conditions found in Table 2 [234, 281]. The process of BSA fibrillation deviates from typical, lacking a defined lag phase similar to acylphosphatase, a small mixed α/β protein, indicating that the nucleation step may be highly favorable [181, 234]. Also BSA aggregates deviated from the robust resistance to protease typically seen in amyloid fibrils comprised of A β peptide, lysozyme or β 2-microglobulin [234, 282-284]. BSA aggregation under both conditions supports the general multistage fibrillation mechanism where under a specific set of conditions BSA becomes destabilized and partially or fully unfolded, forming a molten globule-like state under low pH and denature state under physiological pH, respectively. This results in a decrease in α -helical content, giving rise to an increase in β -sheet content [234, 281, 285]. The aggregation rate of BSA can be modulated by the addition of NaCl which appears to prevent local monomer repulsions and accelerate aggregation [281]. The propensity of BSA to readily form amyloid aggregates appears to be increased by the ability of BSA to

form dimers [234]. It appears that a single unpaired cysteine at position 34 facilitates the aggregation process of BSA [234]. It is also interesting to discuss a temporary functional stability that may indicate an evolutionary advantage for BSA to avoid amyloid formation. Binding of BSA to a ligand inhibits the formation of amyloid-like fibrils by stabilizing the native state [234, 286]. It is clear that aggregation into ordered amyloid fibrils is not limited to single domain α -helical proteins and there are conditions that allow for rapid formation of amyloid fibrils.

Despite the significant advancements through the study of amyloid fibril formation and structure there is still much to be determined. One of the largest problems associated with furthering our understanding is obtaining the detailed atomic structure of amyloid fibrils. The heterogeneous size and insolubility of fibrils make them difficult to crystallize or study via solution-state NMR. However, advances in solid-state NMR methodology, spectroscopic techniques and the formation of microcrystals have paved the way to eventually obtaining high-resolution amyloid fibril structures. Of particular importance is resolving the structures of on pathway soluble oligomeric species as they are now believed to play a critical role in cytotoxicity in some fibril associated diseases such as Alzheimer's [287]. Many fascinating and essential questions still drive current research and are likely to provide greater understanding of the balancing act between protein folding, misfolding and aggregation as the field continues to progress into the future.

METHODS TO STUDY PROTEIN MISFOLDING: THIOFLAVIN T AND CONGO RED BINDING FLUORESCENCE

In order to diagnose amyloid related diseases and study the mechanism of fibrillation there is an important need for molecular probes. The β -sheet rich morphology of amyloid fibrils is susceptible to the binding of histological dyes for biomedical assays. The most commonly used dyes used for the detection of amyloid fibrils are ThT and CR (Figure 20). It is not clearly understood how molecular probes bind the amyloid structure because high-resolution techniques are not possible due to the insolubility and heterogeneous nature of amyloid fibrils. ThT is a fluorescent dye that shows a large fluorescence enhancement with excitation and emission maxima at about 450 and 480 nm respectively, once bound to amyloid fibrils; and is used as both a visualization and quantification method [288, 289]. It is still quite controversial as to how the ThT binding to amyloid fibril structures causes a photophysical enhancement in the fluorescence [289]. ThT fluorescence microscopy is used to investigate fibrillation *in vitro* and as an amyloid diagnosis in some clinical investigations of tissue sections [288, 290, 291]. Other spectroscopic uses of ThT include direct observation of fibrillation by internal reflection and anisotropy fluorescence [292-294]. Amyloid fibril quantification is possible using ThT due to the proportionality of fluorescence intensity to the fibrillar weight concentration for a given protein under a specific set of conditions [288, 295]. However, it should be noted that several factors may affect the spectral fluorescence intensity when used under different conditions. These include the specific protein used, fibril morphology (high viscosity), ThT concentration, pH, ionic strength and the recognition that ThT may sometimes bind to tissues themselves [294, 296-302]. Unfortunately, it

appears that ThT may actually promote amyloid aggregation to a small degree thus *in vitro* uses are becoming more limited [292, 294, 303]. In addition, there is evidence that suggests that ThT is also unable to interact with the protein fibrils and potentially neurotoxic prefibrillar species such as oligomers [287, 304-306].

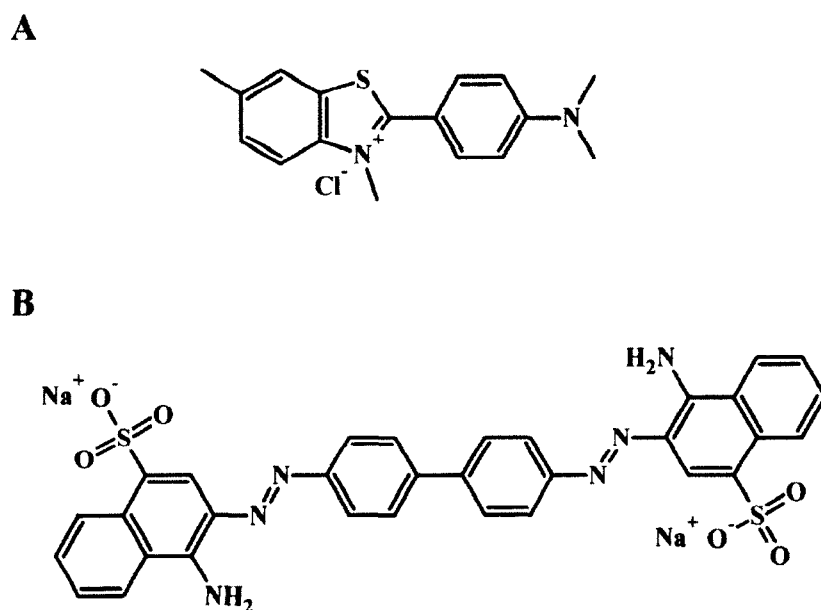


Figure 20. Chemical structure of two commonly used molecular probes for amyloid fibril detection. Structures shown are (A) ThT and (B) CR.

For almost a century CR has been used as a histological staining dye for to diagnose amyloid related diseases [307, 308]. It is the most common diagnostic test for detection and identification of amyloid aggregates in tissues in combination with polarized light microscopy [307, 308]. The binding of CR to the amyloid structure results in an apple-green birefringence when examined by polarization microscopy (Figure 21). In addition, using UV-Vis absorption spectroscopy, there is a characteristic shift in the

absorbance maxima of CR from 490 nm to about 540 nm when bound to the amyloid form [292, 309-312]. CR absorption has been used as a quantification method for insulin and amyloid β aggregates *in vitro*, although the binding of CR is limited by a lack of sensitivity at low amyloid concentrations [313, 314]. CR absorption analysis is also poorly suited to *in situ* detection because it has been reported to interfere in fibrillation by either inhibiting or enhancing the process [311, 312, 315-318].

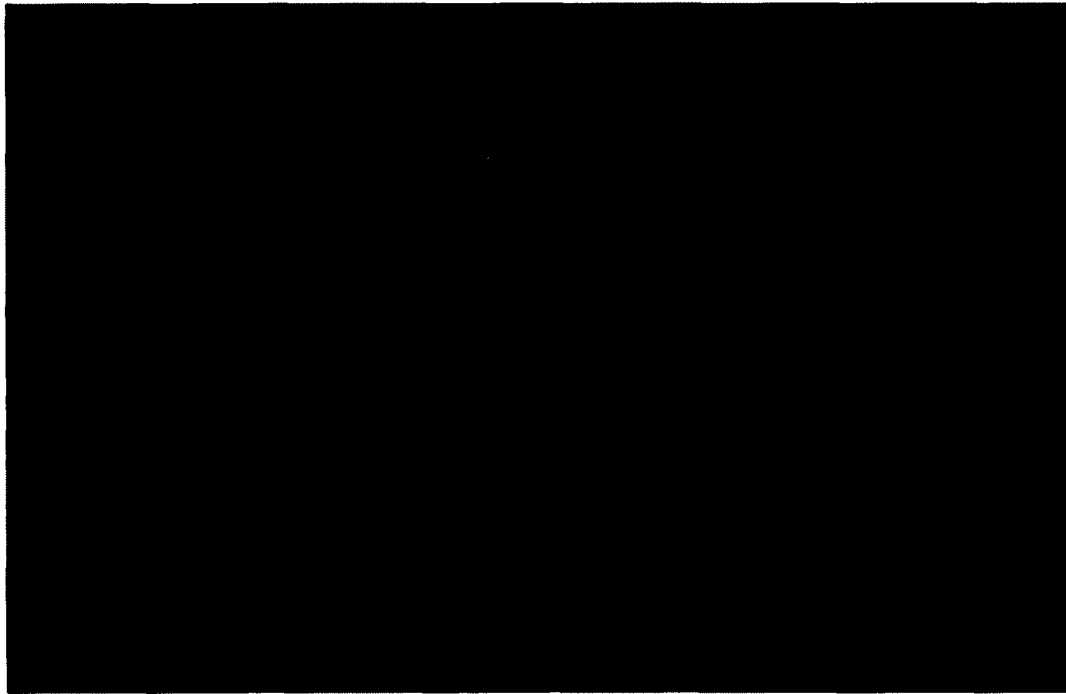


Figure 21. Congo red apple-green birefringence for the rare laryngeal amyloidosis. Amyloid fibrils in this image involve a protein immunologically identical to the variable region of the light chain fragment of immunoglobulin. This image was reproduced from [319].

Both ThT and CR have been proposed to bind in different molecular forms, such as monomeric [311, 320-328], dimeric [324, 329] or micellar [330-333]. The orientation of ThT and CR in the binding pocket of fibrillar structures has been investigated via molecular dynamics simulations indicating the possibility of either parallel or perpendicular binding to the β -strands in the β -sheet of the fibrillar structure [334]. Parallel binding would indicate intercalation between the β -strands however, X-ray fiber diffraction experiments show that the interstrand spacing of 4.8 Å is unchanged upon binding [288]. A change in the internal β -strand distance would be expected due to the ThT binding. Binding of ThT or CR is more commonly believed to orient and bind in a perpendicular fashion to the β -strands intercalating between the β -sheets parallel to the long axis of the fibril strand [320]. X-ray fiber diffraction revealed a change in the inter-sheet distance from 11 Å to 16 Å when bound to ThT [288]. It is essential to further understand the binding modes of these molecular probes and their derivatives for a number of reasons. Complete understanding of how these aromatic structures bind to the amyloid structure will also inform the design of nontoxic small molecule inhibitors as there is much evidence that derivatives of ThT and CR have been shown to inhibit fibril formation [315, 335-337]. We can also improve interpretation of the amyloid signature and enhance our understanding of the amyloid structure.

METHODS TO STUDY PROTEIN MISFOLDING: TRANSMISSION ELECTRON MICROSCOPY

In the study of amyloid fibril structure and morphology one of the most crucial techniques is the application of TEM. TEM opens the door to understanding the polymorphisms associated with amyloid fibrils. It is also one of the primary methods for investigating amyloids formation by providing snapshots of the gross fibrillation process. The ability to resolve and analyze structural details on the nanoscale provides significant insights into the area of protein misfolding. TEM has advanced tremendously in the last 50 years in their design, preparation protocols and analysis capabilities [338]. The TEM is an advanced imaging system that utilizes electrons passed through a sample to produce contrast, similar to how objects in front of a light source produce a shadow. TEM can image the details of a sample with magnifications of up to 10^6 times and with a resolution of less than a tenth of a nanometer [338]. The TEM is composed of four integrated systems: the illumination system, the sample manipulation system, imaging system and the vacuum system (Figure 22). The illumination system is at the top of the microscope and produces the uniform electron beam of homogeneous energy that is directed onto the sample [338]. There are two predominate types of electron emitters: the typical electron gun with a tungsten wire or lanthanum hexaboride cathode and the field emission gun using a tungsten crystal [338]. The sample manipulation is quite a sophisticated system in that the sample stage must be able to move smoothly in the x and y plane over a distance of 250 mm in steps as fine as 10 nm. In addition to the ultrafine movement control, it must also be able to maintain a stable stationary position varying no more than 0.1 nm for a minimum of 3 seconds for imaging purposes [338]. The imaging system uses

electromagnetic lenses to form, focus and magnify the sample image onto the viewing screen or imaging system. Of these electromagnetic lenses the objective lens is the single most important lens in the TEM as it is responsible for forming and focusing the initial image [338]. The viewing system at the base of the TEM uses electron phosphors to image the electrons that were not scattered when passing through the sample which shows up bright while places where electrons were scattered will show up darker [338]. The final integrated component is the vacuum system which is fundamental in its operation because the illuminating electrons would be easily deflected by collisions with any gas molecules that were present [338].

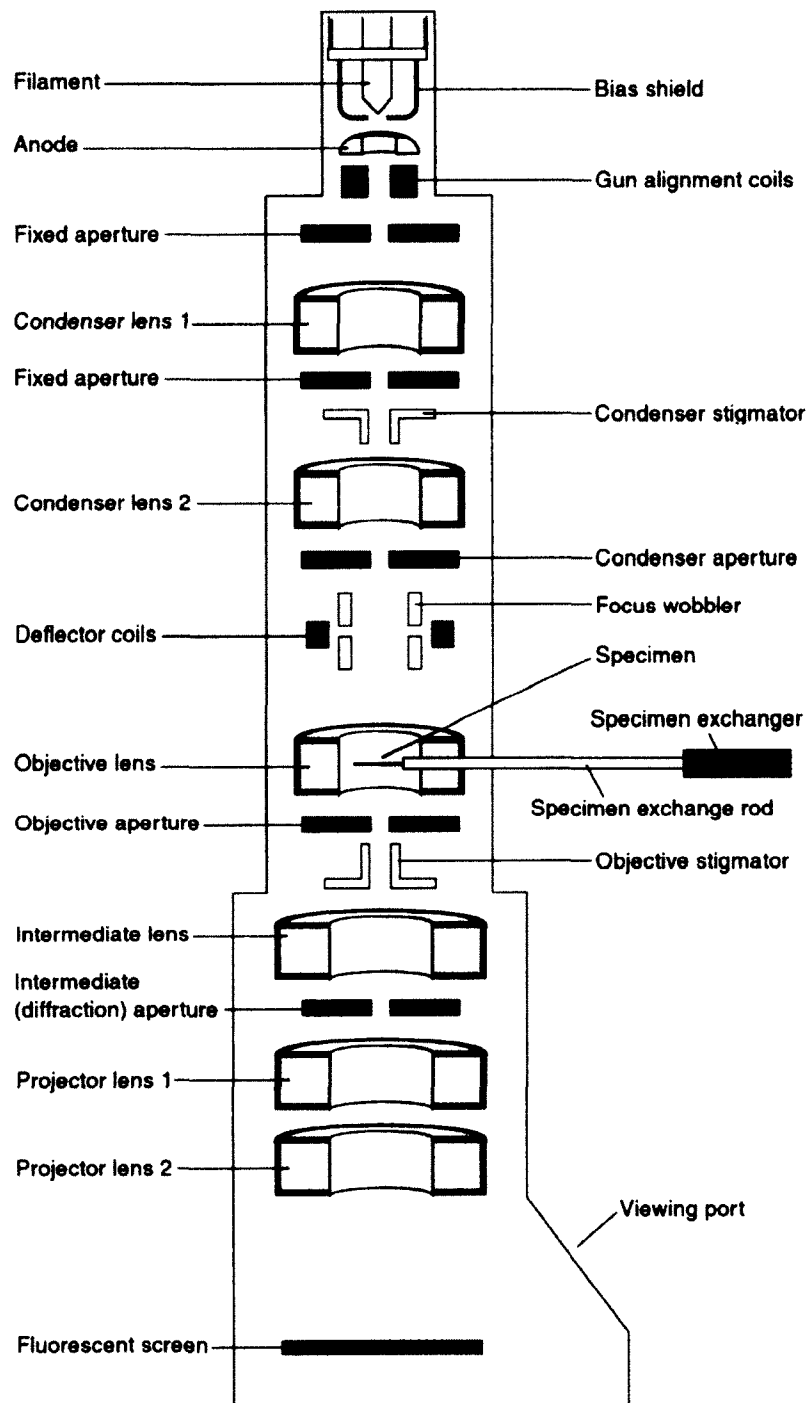


Figure 22. Schematic of the column design of a standard high-resolution TEM. The TEM is generally divided into the illumination system (filament, bias shield, anode and gun alignment coils), the sample manipulation system (specimen exchanger rod) and imaging system (screen, lens and apertures). Image reproduced from [338].

When imaging biological samples it is critically important to consider the appropriate sample preparation required for high resolution imaging of a sample. When microbiological samples are used preparation involves chemical fixation, sectioning and staining. Chemical fixation is a preservation method for stabilizing tissues and cells in as close to a “life-like” state as possible for imaging without apoptotic artifacts. Formaldehyde and glutaraldehyde are a few examples of chemical fixatives that are commonly used. In TEM the electrons must pass through the sample in order to be visualized. This can be problematic as most samples of interest are too thick for electrons to penetrate. Thus, ultramicrotomy is employed using a diamond tipped knife to cut cross-sections of a sample ranging from 60 to 80 nm which is thin enough to allow electrons to pass. In studying the amyloid fibril structure *in vitro*, chemical fixation and sectioning is not necessary to visualize them by TEM. However, where sample preparation plays a critical role for the study of amyloid fibrils is in the staining process. In the preparation of most samples TEM grids are used as a scaffold for the sample to be stained and imaged. Because all biological material is composed of predominantly carbon, nitrogen and oxygen, imaging by TEM is extremely difficult as electrons are not deflected or scattered by these atoms. To overcome this heavy metal stains are used to coat and/or stain samples which are ideal for deflecting electrons due to their heavy nuclei. In studying amyloid fibrils some of the common stains used, i.e. uranyl acetate and lead citrate, can lead to a significant contrast increase. Due to the heterogeneity and solubility issues of amyloid fibrils in solution, TEM serves as one of the best methods for investigating the structure of these polymers. In an extensive investigation of the polymorphic amyloid structure of the amyloid β -peptide (1-40), TEM was critical in

showing the unique twists formed during fibrillation [339]. In a more recent study of hen egg white lysozyme fibrillation treated with functionalized gold nanoparticles TEM was used as the primary high-resolution imaging method to observe the resulting effect [340]. Interestingly, a curcumin-functionalized gold nanoparticles was shown to not only inhibit amyloid fibril formation but also were able to dissolve the fibril structure [340]. The value of TEM in the investigation of amyloid fibrils and protein misfolding is immeasurable.

METHODS TO STUDY PROTEIN MISFOLDING: ATOMIC FORCE MICROSCOPY

Atomic force microscopy (AFM) was developed in 1986 by Gerd Binnig and is a surface technique for studying surface topologies with atomic-level resolution [341, 342]. AFM was developed as an offshoot of scanning tunneling microscopy (STM) in which the sharp tunneling tip is replaced by a force-sensing cantilever. In STM atomic resolution is achieved by bringing an electrically conductive tunneling tip to within angstrom distance from the surface and measuring the tunneling current that is induced [343]. Distance of the tip and the current induced are inversely proportional allowing for topography imaging of a surface with atomic-level resolution. However, the drawback to this method is the material needs to be electrically conductive and placed in an ultra-high vacuum [343]. With those restrictions imaging biomolecules like proteins or amyloid fibrils is next to impossible. By replacing the tunneling tip with a force-sensing cantilever, the requirement for electrical conductive material is eliminated and a more diverse set of sample types can be visualized. In addition, these samples could now be visual under ambient conditions with high-level resolution. The AFM also retained the ability to visualize samples up to the atomic-level using ultra-high vacuums, previously achieved by STM [343]. Later, AFM was further developed by the introduction of the dynamic vibrating probe which provided increased versatility and resolution for material topometry [344-346]. Because of its versatility, dynamic AFM techniques are becoming the predominate method for imaging DNA, proteins and polymers in both liquid and air mediums [347-350]. A schematic of the fundamental design of AFM is shown in Figure 23A. Amplitude (AM) and frequency (FM) modulation AFM are the two major modes

used in force microscopy [344, 345, 351]. The force-sensing cantilever in dynamic AFM has a chemically or mechanically etched tip on a stiff microlever. AM-AFM (also known as tapping-mode) is a direct contact method for imaging surface dynamics [346]. The tip is excited at its resonance frequency and the oscillation amplitude is used to measure the surface dimensions [346]. Once in contact with the sample, changes in the z-axis increases or decreases the oscillation amplitude and are visualized as changes in surface topology. In addition, information about the material properties can be gained by looking at the phase shift between the driving force and tip oscillation [346]. An example of AM-AFM image on a protein structure is shown in Figure 23B. On the other hand, in FM-AFM (also known as non-contact mode) the cantilever oscillation amplitude is fixed and the image contrast comes from the forces between the sample and tip [346]. Attractive and/or repulsive forces of the cantilever to the surface modulated the resonance frequency of the free lever providing differentiation of the surface dimensions. Typically, FM-AFM is used in an ultra-high vacuum for atomic-level resolution whereas AM-AFM is used predominantly in air or in liquid samples for nanoscale topography [346].

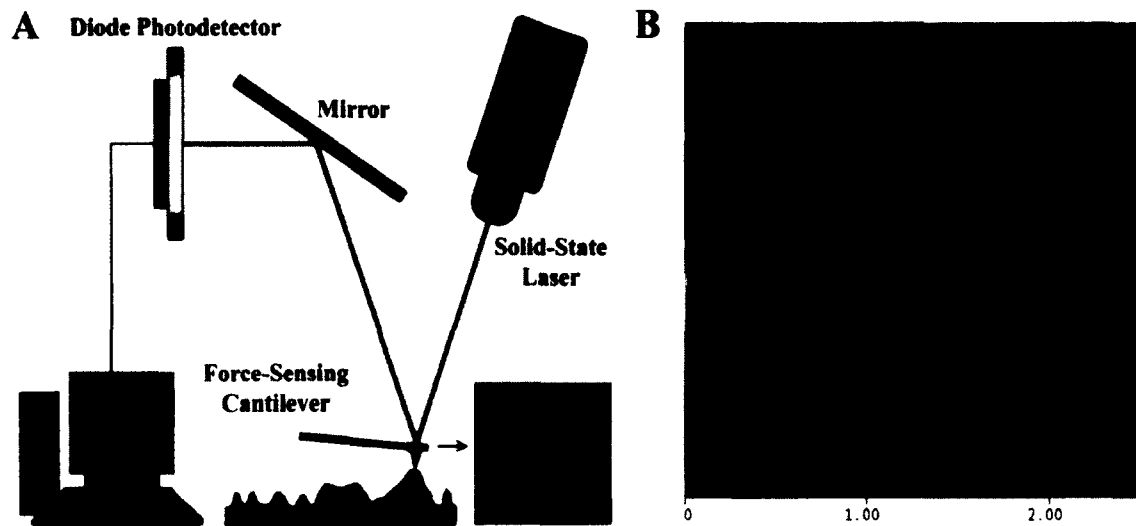


Figure 23. A fundamental schematic of the design of afm and example image obtained by AM-AFM. (A) Image of the cantilever tip is taken from Nanotec (<http://www.team-nanotec.de/index.cfm>). **(B)** AM-AFM image taken of β -synuclein fibrils using a Veeco DiNanoscope 3D AFM. The x-axis is in μm and the z-axis scale bar is from 0-50 nm.

HIGH-RESOLUTION ANALYSIS OF α -HELICAL PROTEIN FIBRILS AND THEIR POLYMORPHISMS

The amyloid fibril structure is composed of regular repeating highly ordered macrostructures composed of a stacked array of misfolded cross β -sheet proteins along the fiber axis [88, 177]. The preliminary association monomeric cross β -sheet proteins make up the protofibrillar species, typically 2-5 nm in diameter, which further associate resulting in formation of single fibril strands ranging from 5-13 nm in diameter depending of the conditions of fibrillation [88, 352]. There is a clear indication that the fibrillation process can occur via multiple pathways (Figure 24) [353]. Detailed investigations of the fibril morphology is typically done using high-resolution electron microscopy and atomic force microscopy imaging, which have revealed repetitively twisted cross β -sheet filaments of variable morphology and length (Figure 25) [339]. Polymorphisms of amyloid fibrils appear to be diverse and not dependent on the amino acid sequence. Here we will compare the unique polymorphic structures of amyloid fibrils obtained from all α -helical proteins. Because transition of all α -helical proteins into the β -sheet rich amyloid fibril form represents the most extreme conversion this section will focus on the all α -helical protein fibril morphology. Understanding the subtle differences in morphology may hold the key to developing pharmaceuticals that specifically target toxic amyloid structures.

Nucleation and Growth

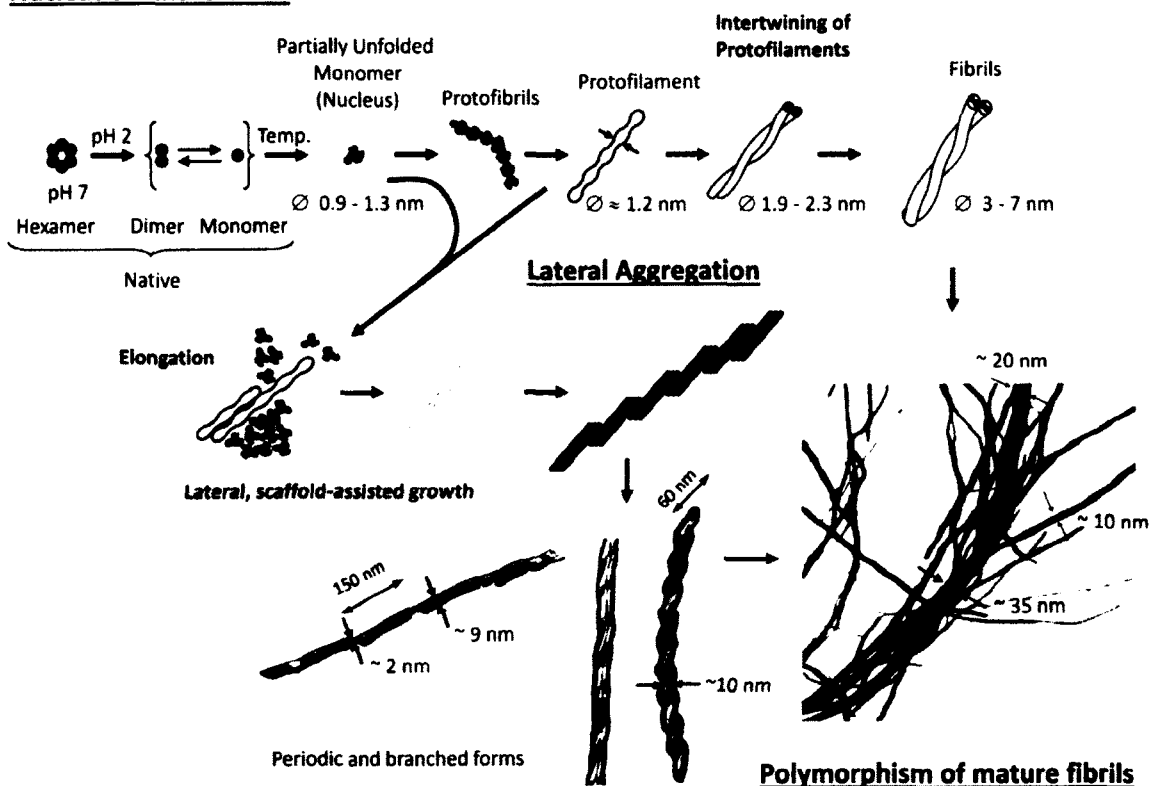


Figure 24. Schematic of the proposed polymorphic aggregation mechanism of insulin. In the nucleation and growth phase the native-state transitions to a fibril form. These associate through lateral aggregation into a higher-level and more complex state. Lastly, these lateral forms can have variable morphologies which are unique polymorphisms specific to mature fibril types such as insulin. Reproduced from reference [229].

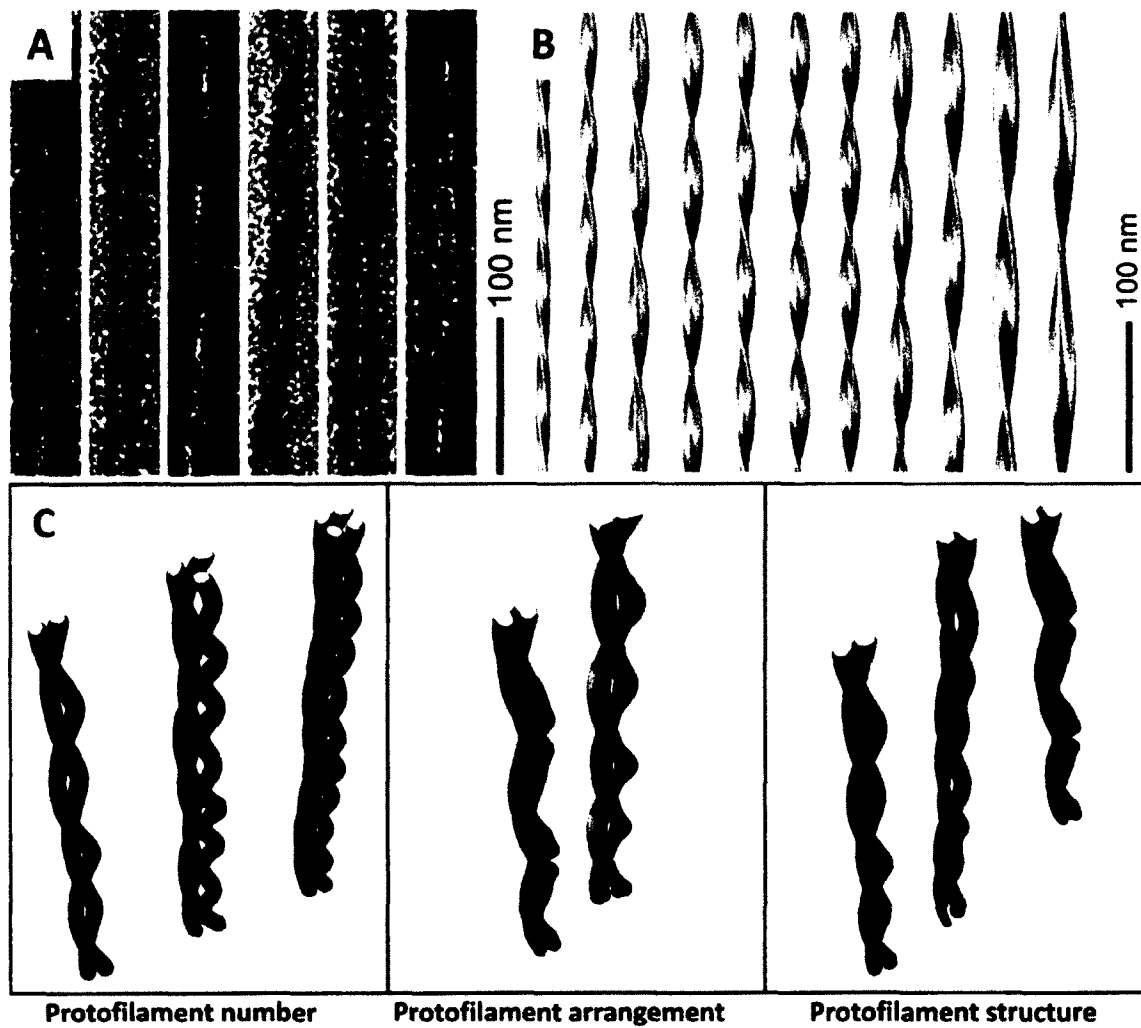


Figure 25. Imaging and modeling of a non-helical protein and associated amyloid fibril morphologies. (A) Transmission electron microscopy of types of polymorphic amyloid fibril structures. (B & C) Models of several organizational patterns for the association of protofibrillar species into various morphologies. Reproduced from reference [339].

Apolipoprotein A-I

Apolipoprotein A-I presents with what appears to be two fibril morphologies, one is that of a twisted ribbon with irregular twisting pattern and the other is a similarly twisted circular fibril loop (Figure 26) [253]. The ribbon-like structure may be due to lateral associations of fibril strands producing a flat ribbon of 11 nm in width [253]. The circular morphology indicates that the fibrils formed by apolipoprotein A-I may be dynamic in flexibility allowing the fibril ends to find one another. A similar morphology is seen in the fibrillation of apolipoprotein CII [354, 355]. Interestingly, X-ray diffraction indicates that the formation of amyloid fibrils from a specific Leu174Ser mutant of apolipoprotein A-I presented with a heterologous substructure of crossed- β and coiled-coil helical morphologies [356].

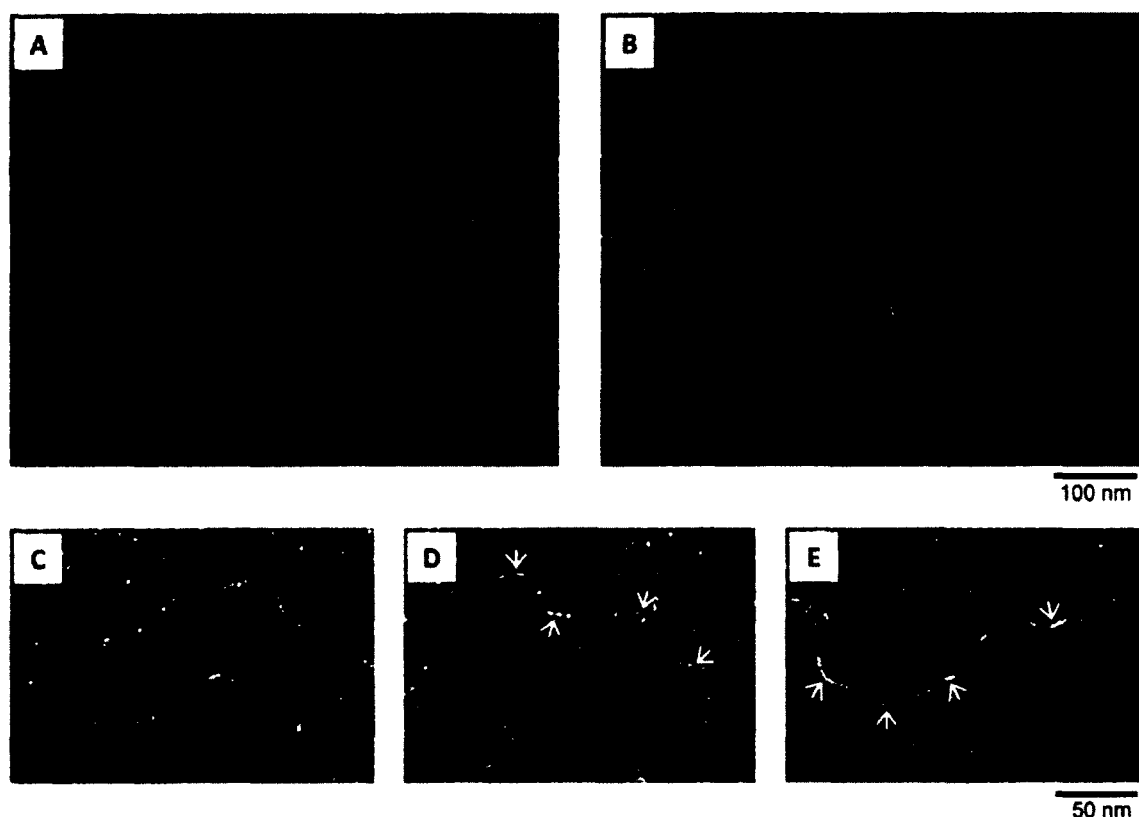


Figure 26. Circular polymorphism of apolipoprotein A-I amyloid fibrils. Electron microscopy imaging of (A) circular and (B) filamentous apolipoprotein A-I amyloid fibrils. Close up of a (C) circular and (D & E) linear form of apolipoprotein A-I. Reproduced from reference [253].

Insulin

The fibrillation of insulin is one of the most well studied aggregation processes and demonstrates a highly polymorphic nature under various conditions for fibrillation. It has been shown that fibril morphology of insulin is altered by differences in fibrillation conditions [229, 353]. In an extensive investigation of insulin fibrils using high resolution AFM we can clearly see the polymorphic assembly of insulin amyloid fibrils (Figure 27)

[229]. The aggregation of insulin fibrils seems to only rely on the formation of non-covalent interactions under destabilizing conditions mentioned previously [229]. Higher ordered fibril associations of insulin fibril strands occur rapidly within 30 minutes of incubation at 60 or 70 °C forming branched, twisted and lateral associations of amyloid fibril structures [229]. According to AFM height distributions of protofibrillar forms of insulin, the association of insulin into amyloid protofibrils produces aggregates that are 1.1 nm in diameter [229]. Fully assembled single strands of insulin fibrils shown by TEM and AFM are 1-1.2 nm in diameter [229, 357]. Strands of insulin protofibrils assemble into intertwisted and laterally associated early fibrillation strands that are ~4.4 nm in diameter when incubated at 60 °C and 2.9 or 6.5 nm at 70 °C with a regular repeating helical rotation [229]. In one case, twisting of the insulin fibrils produced braid like threads of 2.2 nm indicating two intertwined strands [261]. Maturation of insulin fibrils results in an increase in polymorphic structures currently classified as parallel tubular fibers, twisted ribbon-like structures, rod bundles and ropelike textures with increasing sizes ranging from 4-9 nm in diameter (Figure 27A) [229]. Mature insulin fibrils are able to assemble into amyloid-like fibrils via a multiple pathway process, which include a hierarchy of strand intertwisting or assemble via lateral associations of twisted threads, producing different structural polymorphisms (Figure 24).

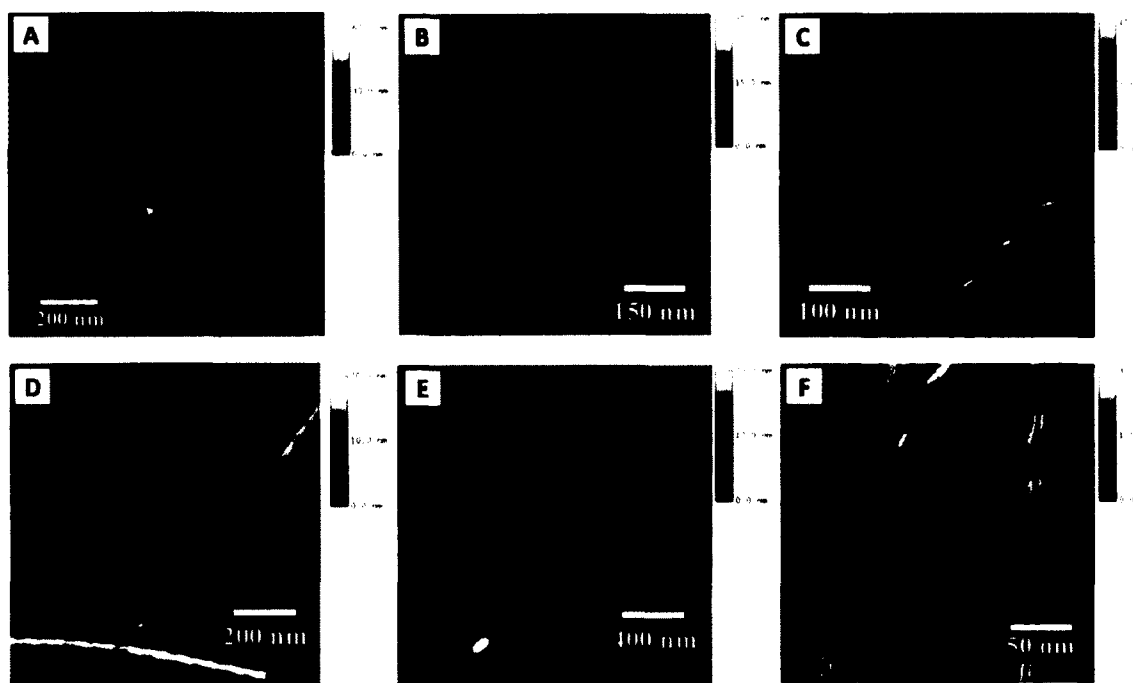


Figure 27. AFM imaging of insulin fibrils and their associated polymorphisms. The various morphologies of insulin include (A) twisted morphologies with different fibril diameters (arrows), (B) laterally connected double strand fibers, (C) highly-twisted with overlapping morphologies and (D & E) chain-like stacking of preformed amyloid subunits. (F) The interior canal of a single insulin fibril thread is visible by AFM. Reproduced from reference [229].

Lung surfactant protein C

LSP-C fibrils isolated from patients with pulmonary alveolar proteinosis and imaged with TEM show long amyloid fibrils of various twisted morphologies however higher resolution imaging is necessary for clearer determinations to be made (Figure 28A) [230]. However, to our knowledge, high resolution polymorphisms of *in vitro* amyloid-like fibril formation of LSP-C have not been assessed.

Prolactin

Amyloid fibrils formed from prolactin (Figure 28B) and other hormone peptides except for glucagon and gastric inhibitory polypeptide appear to be homogeneous in morphology and further higher resolution investigations into the possibility of polymorphisms needs to be further evaluated (Figure 28C-L). Glucagon is a unique case in that at equilibrium its native state appears to include both unfolded and α -helical species in solution, however it has been shown to form three distinct polymorphic structures similar to those discussed in insulin [358]. Gastric inhibitory polypeptide appears to have a regular twisting pattern indicating the presence of possible alternate morphologies. However, further investigation is also required to assess alternative polymorphic structures.

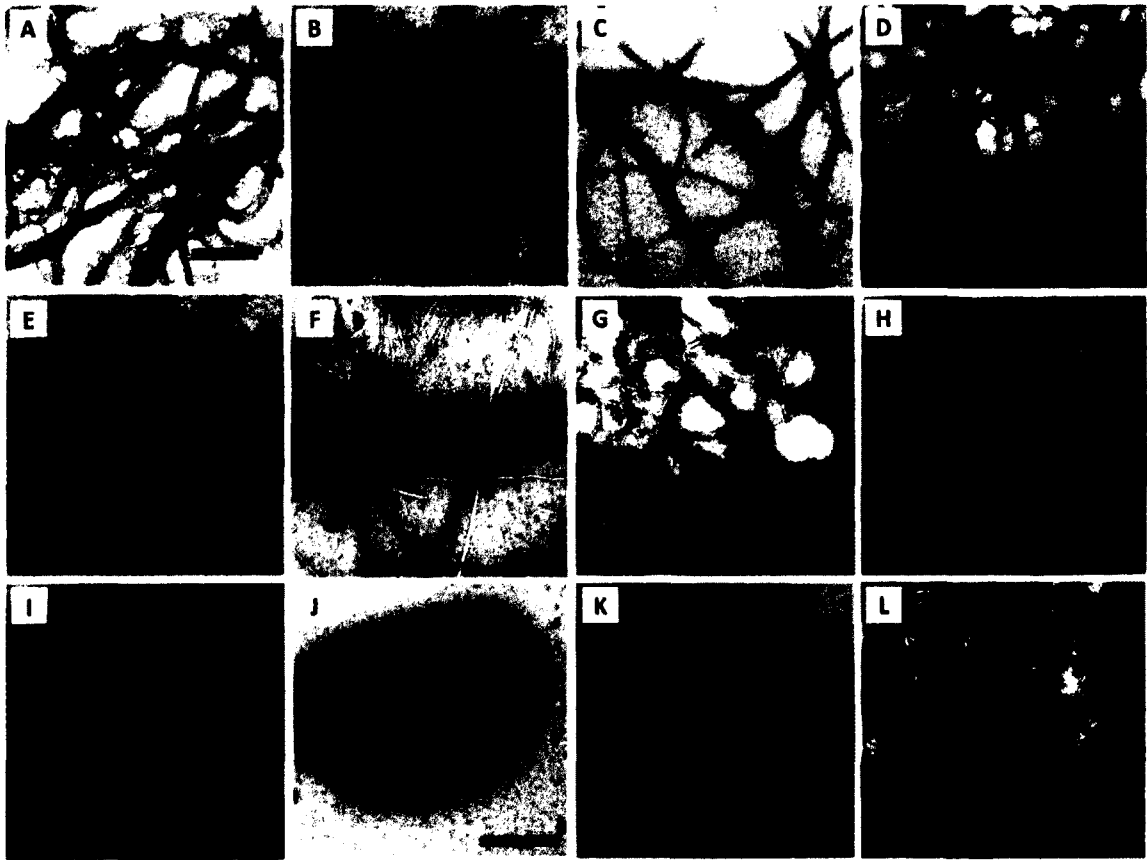


Figure 28. Amyloid fibril imaging of other α -helical proteins associated with a disease state or functional amyloid form. Electron microscopy imaging of (A) lung surfactant protein C, (B) prolactin and helical hormones; (C) glucagon, (D) corticotropin releasing factor, (E) exendin-4, (F) gastric inhibitory polypeptide, (G & H) glucagon-like peptide 1 & 2, (I) neuromedin K, (J) neuropeptide Y and (K & L) urocortin II & III. Scale bars indicate 500 nm distances. Reproduced from references [195, 230].

Bovine Serum Albumin

There are indications that under different conditions the aggregation process of BSA may follow different pathways and form morphologically distinct fibrils (Figure 29G & F). BSA fibrils formed at 65 °C only bound ThT dye and CR, forming curly fibrils which were ~10 nm in diameter [234]. At 60 °C ThT indicated fibrillation occurring whereas CR was negative, which indicates that there may be a distinction in fibrillar structures at these two temperatures [234]. However, more detailed investigations of BSA aggregation are required as the curly fibril structures indicate that the conditions for fibrillation may not be optimal. Interestingly, BSA fibrillation at a more basic pH produces aggregates that are 2-4 nm in diameter however no distinction between the possibility of polymorphisms were assessed [235]. Under acid pH conditions BSA amyloid fibrils appeared more morphologically similar to typically seen thread like filaments found in A β peptide and lysozyme indicating that these conditions may be more optimal and visual examination of TEM show fibrils of significantly different sizes indicating the presence of differing morphologies [281].

Apomyoglobin

Morphology of WT apomyoglobin mature fibrils (Figure 29E) appears to be indistinguishable from fibrils formed in disease state amyloids; whereas fibrils formed from Trp7Phe/Trp14Phe apomyoglobin exhibited branching after long periods of incubation (Figure 29F) [182, 276]. Visual analysis of TEM of WT apomyoglobin fibrils show two types of fibril structures, curly and rod-like, having approximate diameters of 10 and 14 nm respectively [276]. Interestingly, fibrils formed from the Trp7Phe/Trp14Phe apomyoglobin mutant formed fibrils that were on average much

larger in size, 20-50 nm in diameter. Conversely, fibrillar aggregates formed by mutants of residue Val10 appeared twisted and smaller in size having a diameter of ~7-10 nm [279]. In these cases TEM data suggest that there may be polymorphisms that have yet to be distinguished in the current research of these proteins.

Cytochrome c_{552}

Based on electron microscopy imaging of cytochrome c_{552} amyloid fibrils have a typical 6-13 nm diameter with variable lengths indicating the presence of possible polymorphic structures (Figure 29C) [236]. Electron microscopy imaging of the fibrils produced from the bovine homolog of cytochrome c (Figure 29D) showed similar fibrous thread-like fibrils. However, no indications were provided as to the morphology or the possible presence of polymorphic structures [233]. Visual analysis of the TEM image indicated large fibril structures ~15-30 nm in diameter [233]. Further, detailed investigations would be required to identify polymorphic assemblies of cytochrome c_{552} and bovine cytochrome c .

Apoptotic Protease Activating Factor-1 Caspase Activation and Recruitment Domain

The process of Apaf-1 CARD fibrillation is believed to proceed from a destabilized molten globule-like intermediate into precursor aggregates which assemble into protofibrils and elongated fibrils [232]. Analysis by AFM shows that the precursor aggregates are ~2.1 nm in diameter indicating they may not be fully formed into the crossed β -sheet structure indicative of amyloid fibrils [232]. Interestingly, protofibrillar and elongated fibrils produced by the Apaf-1 CARD protein are heterogeneous in length, but homogeneous in morphology with a single width distribution of ~2.6 nm indicating that there is no intertwining of fibrillar strands (Figure 29A) [232]. This is interesting as it

does not fit the structural variety typically seen in amyloid fibril morphology having larger twisted structures containing multiple strands. These sizes are much smaller than what is found in a classical amyloid fibril, which may indicate that the conditions for fibrillation in this case may not be optimal for aggregation.

Fas-associated Death Domain

Interestingly, agitation seems to be pivotal in tipping the balance between ordered and unordered aggregation conditions for Fadd-DD fibrillation [183]. The Fadd-DD fibrils that formed optimally are rod-like filaments that are uniquely shorter than fibrils generally seen (Figure 29B) [183]. Analysis of fibrillation using TEM indicates that Fadd-DD fibril length can be modulated by differences in agitation conditions [183]. High agitation (180 rpm) results in a predominance of morphologically shorter Fadd-DD amyloid fibrils, approximately 0.5 μm in length, whereas lower agitation (75 rpm) results in approximately a doubling in fibril length to 1.0 μm on average, but is accompanied by the formation of unordered aggregates [183]. Using AFM analysis of morphology of the Fadd-DD aggregation process indicates that protofibrillar species are on average 4 nm in diameter [183]. It appears that association of protofibrillar species results in the formation of single filaments, with similar size to protofibrillar structures, of mature amyloid-like morphology [183]. From the AFM size data analysis of mature fibril aggregates after about 11 days show two types of average size distributions, the second size being 8 nm indicating a possible twisting of two fibril strands [183]. Further investigation of the possible presence of polymorphic structure within Fadd-DD fibrillation needs to be further investigated.

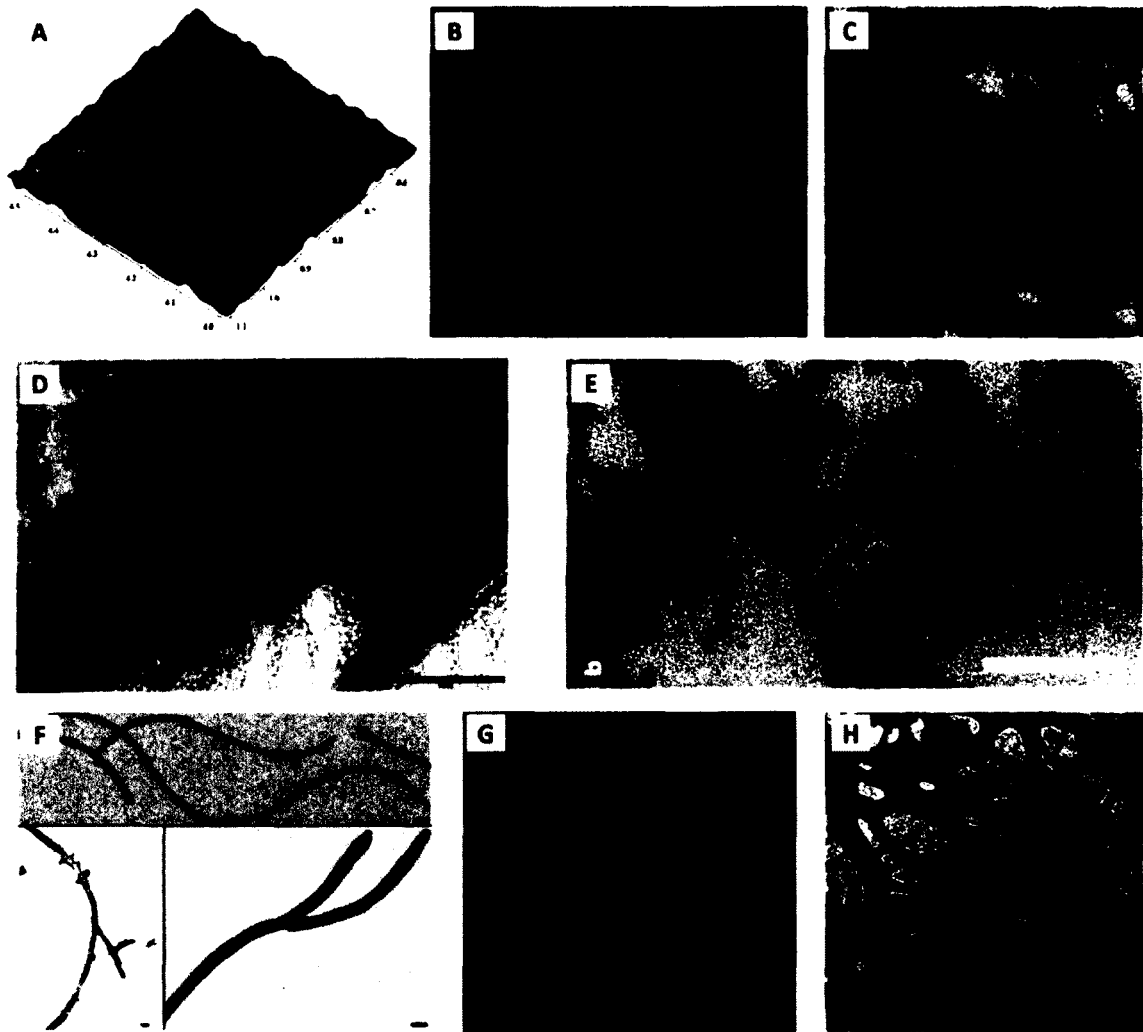


Figure 29. Amyloid-like fibril imaging of α -helical proteins not associated with a disease state or functional amyloid forms. Proteins shown are (A) Apaf-1 CARD, (B) Fadd-DD, (C) cytochrome c_{552} , (D) bovine cytochrome c , (E) myoglobin, (F) myoglobin Trp→Phe variant, (G) bovine serum albumin (pH 7.4), (H) bovine serum albumin (pH 3.0). Reproduced from references [182, 183, 232-236].

RESEARCH GOALS

This dissertation is focused on gaining insight into the folding and misfolding process of the relatively small globular proteins GB1 and Fadd-DD, respectively. As previously discussed protein folding is a complex and dynamic process that is the foundation to the protein folding problem and misfolded forms are associated with a number of disease states. In both cases understanding both protein folding and misfolding are critical to not only progressing our scientific understanding of proteins but can be critical in guiding the design of therapeutic drugs for protein misfolding diseases and ultimately curative methods. This dissertation contains computational and experimental chapters that investigate the protein folding and misfolding problem.

In chapter II, the protein structure of GB1 was investigated from a computational stand point. Evolutionarily conserved features believed to be important to produce the $\alpha+4\beta$ fold of GB1 are revealed. The results of a bioinformatics structural alignment are detailed to determine both positional conservation and amino acid character conservation. In addition, bioinformatics investigations were done to determine a possible evolutionary divergence that separated GB1 from a structurally different 3-helical bundle albumin-binding domain of protein G (GA) shown to have an alternative fold while retaining up to 98% similarity in sequence. Using GB1 and GA as our target, position-specific iterative basic local alignment search tool (PSI-BLAST) examinations are conducted to determine a possible common ancestral protein. Molecular modeling is used to construct a possible structure for the resulting sequence.

In chapter III, we use protein engineering methods such as the PCR and site-direct mutagenesis in order to truly understand how individual amino acids that have been

shown to be conserved in structure and are important in the folding process. Using PCR and site-directed mutagenesis we remove select amino acids and investigate the resulting effect on the transition-state stability. This chapter discusses the work currently completed in collaboration with John Bedford (Graduate Student, ODU) on the site-directed mutagenesis of GB1. It will discuss the residues selected for mutagenesis and the conditions to investigate the transition state stability.

In chapter IV, we investigate the structure, thermodynamic stability and kinetic behavior of GB1. Uniformly and specifically ^{13}C labeled GB1 were synthesized using BL21(DE3) and an *E. coli* auxotroph bacteria, respectively. Both uniform and specifically labeled GB1 are synthesized to ultimately investigate a specific set of long-range tertiary interactions found in the core of GB1 by solution- and solid-state-NMR (ssNMR). Initial ^1H and ^{13}C NMR studies are conducted. In addition, pH-dependent 2D heteronuclear-single quantum coherence (HSQC) experiments were done on the specifically labeled GB1. We also investigated ^{13}C - ^{13}C interactions using DARR ssNMR.

In chapter V, we use Fadd-DD as our model system to explore the hypothesis of Christopher Dobson (Professor, University of Cambridge), which proposes that every protein has the potential to form amyloid fibrils. In this chapter the specific extreme conditions required to transition the all α -helical Greek-key Fadd-DD protein into the amyloid-like fibril form is demonstrated. The fibrillar transition is further studied in detail by CD and high-resolution TEM and AFM. In addition, the extremely narrow pathway of Fadd-DD fibrillation is discussed as it relates to evolution of the protein structure.

CHAPTER II

BIOINFORMATIC ANALYSIS OF THE IMMUNOGLOBULIN-BINDING AND THE ALBUMIN-BINDING DOMAINS OF PROTEIN G: A LOOK INTO A POSSIBLE EVOLUTIONARY ANCESTOR

OVERVIEW

Since the completion of the human genome project the growth in the amount of sequence information available has increased significantly and has amplified the need for bioinformatics techniques. With over 23,000 genes in the human genome alone being able to quickly and accurately compare an unknown protein with respect to all the sequences that are uploaded to the gene bank is a great achievement. Bioinformatics is a multidiscipline field of biology, biochemistry, computer science and mathematics methods, that can complement experimental methods to investigate protein structure and function [36, 359, 360]. Techniques employ sophisticated computer algorithms to compare and analyze protein or gene sequences to a large database of known sequences in order to elucidate complex structural, functional and evolutionary relationships [361]. One of the most common bioinformatics techniques currently used is the PSI-BLAST [362-364]. PSI-BLAST functions by constructing a multiple sequence alignment of the BLAST similarity search output. It then constructs a position-specific scoring matrix sequence and performs a BLAST database search using the matrix sequence as the query [363, 364]. The process may be iterated multiple times as new sequences are found from the search and added to the position-specific scoring matrix sequence construct [363, 364]. Generally, the matched sequences will be a part of an evolutionarily related

superfamily. Resulting sequences and their relationship to the query can be analyzed using a multiple sequence alignment.

A protein superfamily is similar to a family lineage tree in that proteins originate from a common ancestor and will typically share similar sequence, structural and functional features (Figure 30). In terms of sequence similarity it has been determined that proteins that contain >40 % similarity are conserved in function [365]. Whereas, sequences conserved in fold only need to contain <25 % and a significant degree of functional diversity. This method of searching and developing a consensual sequence signature, common structure and/or function is very useful. The potential structure and function of an uncharacterized protein can be guided from that of its relatives whose native structure and function have already been determined. Even though there will be evolutionary drift as a protein diverges from a common ancestral sequence, there will be some sequence and structural features retained as they are responsible for the structural stability and function of the progeny protein [361, 366]. Being able to gain insight into the structure and function of an uncharacterized protein is particularly important as it still requires a significant amount of time to determine the structure and function of a protein.

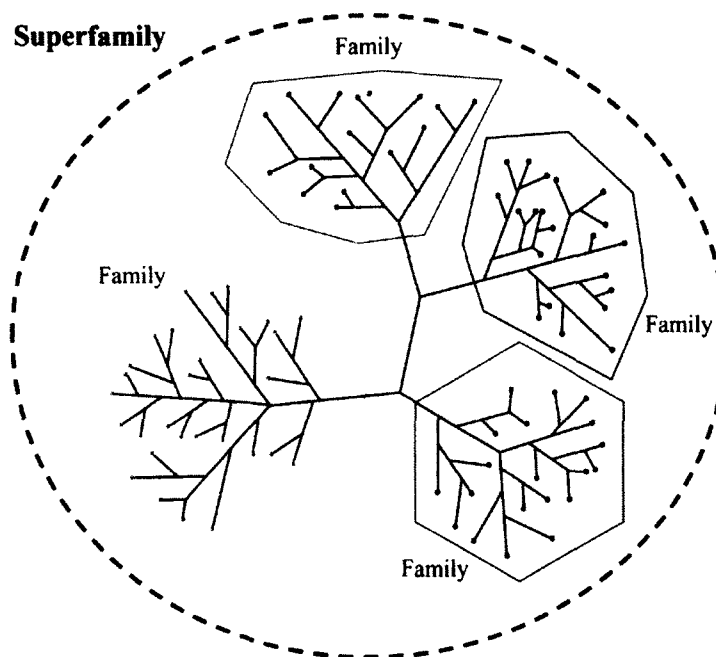


Figure 30. Representative schematic of a superfamily. Individual families of proteins are linked together into a superfamily structure.

We are interested in looking at the evolutionary aspects of proteins capable of switching folds. A protein capable of adopting more than one folded state is now more commonly called a metamorphic protein [367]. The most classic example of proteins with conformational malleability of the polypeptide chain that allow them to switch folds are prions. They are proteins capable of switching from a mostly α -helical fold to a toxic β -rich folded state [368]. In contrast to the toxic prion there are some naturally occurring examples of proteins that can switch folds. These include lymphotactin, mitotic arrest deficient 2 protein and chloride intracellular channel protein 1 which implies that this ability to switch folds may be a more general phenomenon [369-372]. In these types of proteins there is an environmental transition which shifts the equilibrium from one fold

topology to another. These changes can include factors like ionic strength, presence or absence of a ligand and redox state [373]. In other investigations of fold switching involving the Cro family of repressors and the RfaH protein there is a strong indication that evolution may have developed an evolutionary short-cut where a new fold is created from switching an existing structure as opposed to independently developing [374, 375]. The switchable folds have a few commonalities that include, flexible regions and diminished stability required for large conformational changes, a significant degree of uniqueness in the core region to retain the new fold and the development of new functional attributes that stabilize the new fold in order to expand function [376].

In addition to naturally occurring fold switching proteins there have been a number of experimental protein designs used to investigate the nature of the fold switching phenomenon [377-381]. To trace the origins of proteins that switch folds we can use bioinformatics approaches which include searching for related protein sequences and structures and analyzing the nature of their relationship. In this chapter we endeavor to elucidate a possible common ancestral protein between the GB1 and the phage-selected domain-1 (referred to as GA in this dissertation), a construct built from 7 albumin-binding domains (Figure 31) [382]. Both GB1 and GA are 56 residue domains that function in the multidomain *Streptococcus* cell surface protein G but have unique folds [383]. GB1 has the $4\beta+\alpha$ fold and functions in binding immunoglobulin G whereas, GA has a conformational 3-helical bundle and functions in binding albumins found in human serum (Figure 32) [384-387]. GB1 and GA were mutated simultaneously using site-directed mutagenesis in a binary fashion in which only the amino acids of the two proteins are used to increase identity between them [388]. These two proteins have been

shown experimentally to be able to switch folds with a very high (>85%) degree of sequence identity [388-390]. In addition, NMR structures were solved for both GB1 and GA at identities of 88% and 95% and both still retained different folds [389, 390]. The ability of these two proteins to switch topologies with unique single point mutations indicates a possible method by which new functions can evolve.

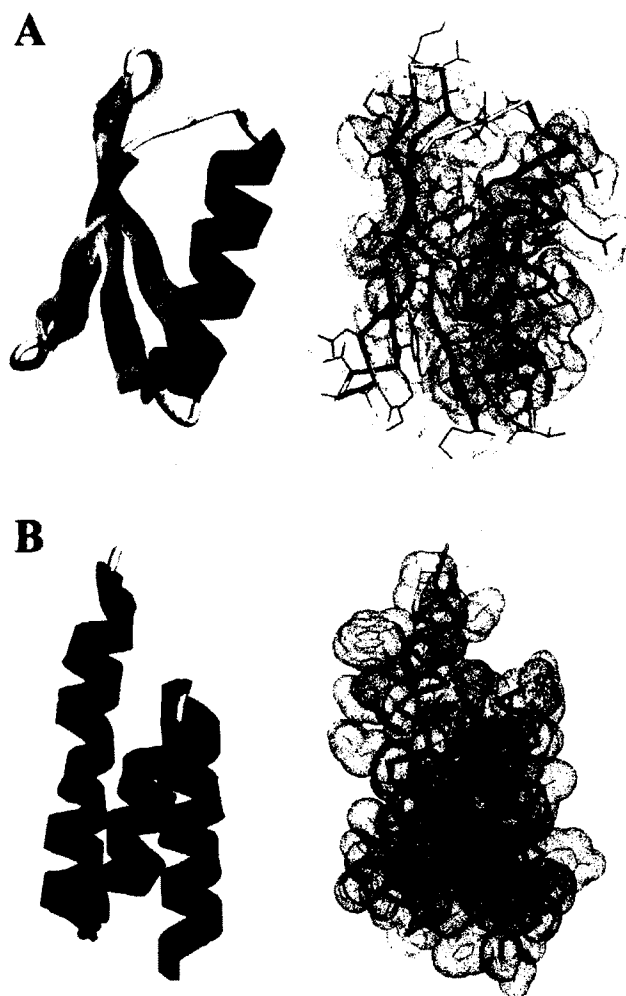


Figure 31. Images of the folds of GB1 and GA. (A) Shows the $4\beta+\alpha$ fold of GB1 and (B) shows the 3-helical bundle fold of GA. On the left are ribbon drawings and the right, the α -carbon backbone with side chains surrounded by Van der Waals radii. The secondary structures α -helices are colored in pink and β -strands colored in yellow. The loop regions, N- and C-terminal are colored grey. All images were created in RasMol (Ver. 2.7.2.1.1).

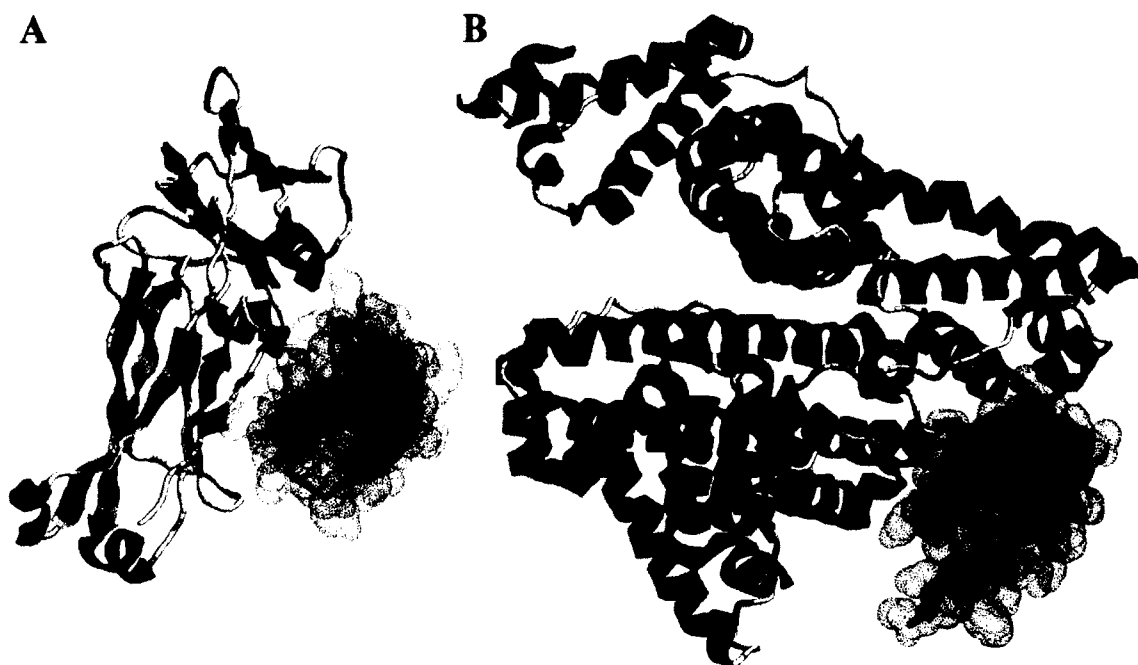


Figure 32. Structural images of ligand bound GB1 and GA. (A) Image of GB1 (red) bound to the Fc region of IgG (PDB code: 1FCC). (B) Image of GA (blue) bound to human serum albumin (PDB code: 1TF0). All images were created in RasMol (Ver. 2.7.2.1.1).

RESULTS AND DISCUSSION

Understanding how proteins evolve is important in elucidating how function and fold can be conserved and will be significant in guiding the engineering of proteins in the future. Fold switching is an important part of protein evolution and GB1 and GA are potentially excellent examples of this phenomenon. In order to truly understand GB1 and GA we need to determine how these two proteins evolved. Did they evolve from one another or are they the result of a larger protein separating into two small ones.

According to the ground breaking work by He *et al.* there are three critical residue positions that result in the populated fold switching between the α -helical bundle of GA and the $4\beta+\alpha$ fold of GB1 (Figure 33). At positions 20, 25 and 45 there is a reversible fold switch from α -helical bundle to the $4\beta+\alpha$ fold by a point mutation: Leu to an Ala, Ile to Thr or Leu to Tyr, respectively [373]. This indicates that they could have evolved from one another, in which the fold switch could have occurred from a point mutation in either GB1 or GA and resulted in a stabilization of the alternate fold over time. However, in my investigation, PSI-BLAST searches of both GB1 and GA did not result in hits of either start protein sequence. This indicates that these two proteins may have evolved from a common ancestral protein rather than the result of an evolutionary fold switch in which a new function is obtained by simply switching the conformation of an existing protein. Rationally, for the folds to switch so easily *in vitro* by a point mutation all of the stabilizing residues for either fold would have to be removed or changed resulting in a significant loss in stability.

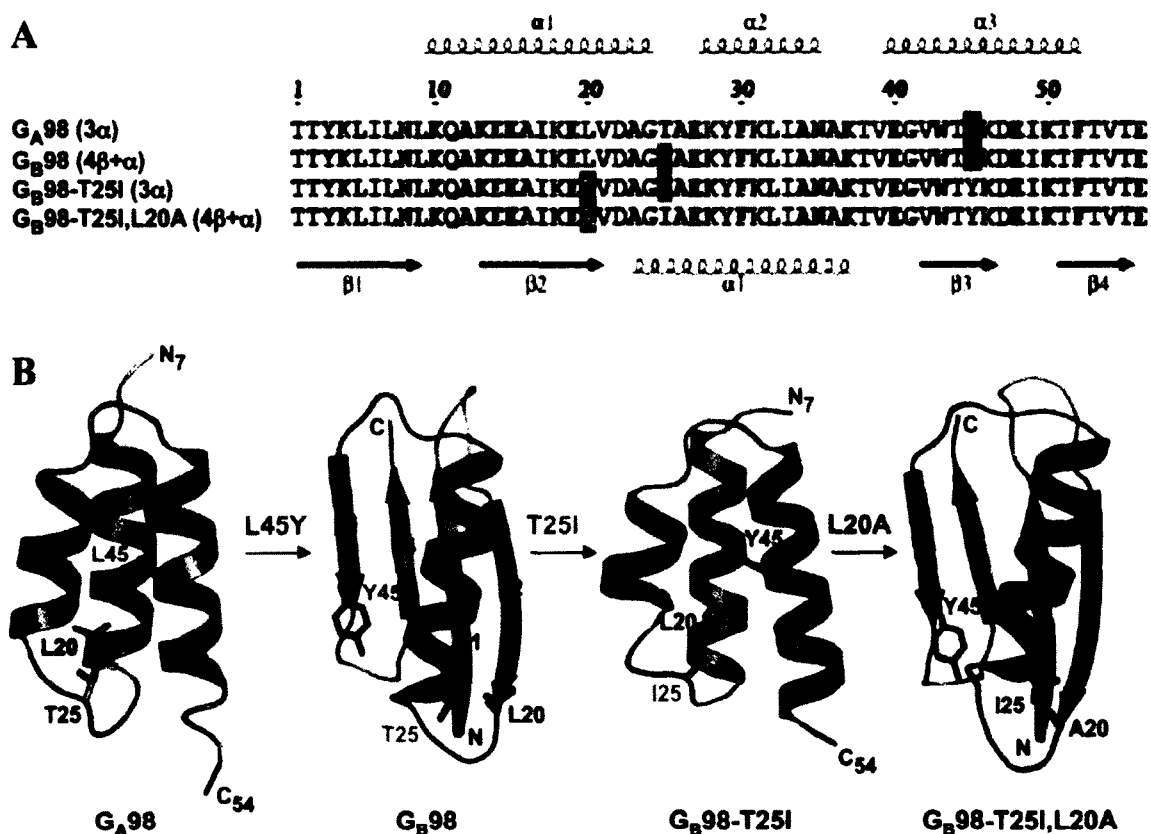


Figure 33. Single amino acid mutations leading to fold switching. (A) Alignment of amino acid sequences for the four proteins with 98 % identity, highlighting the positions at which changes lead to switching between 3 α and 4 β + α folds. (B) Representative structures from the NMR ensembles of G_A98 , G_B98 , G_B98 -T25I, and G_B98 -T25I,L20A. Residues mutated are highlighted. Figure reproduced from [373].

Using a sequence alignment between GB1 and GA show that they have a sequence similarity of 26.9% (Figure 34). This indicates that they are both significantly different from each other. However, investigating further showed that there seems to still be some degree of similarity between GB1 and GA. A sequence alignment of 7 randomly selected small proteins ranging from 50-58 residues with GB1 and GA was constructed

using Multiple Sequence Comparison by Log-Expectation (MUSCLE). The resulting alignment provided an average random sequence identity of about 8 % between the 9 aligned proteins (Figure 35). An alignment of GB1 and GA alone seems to be more related in some aspect as they have almost three times the sequence identity than what is found in a random sampling. Because this was a pairwise comparison of GB1 and GA we needed to see if there was a difference in comparing each sequence with GB1 or GA individually. The comparison resulted in a differing amount of random sequence identity. Thus, a pairwise comparison of GB1 with each of the 7 small proteins using MUSCLE resulted in an increase in average sequence identity to about 12 % (Table 3). This value is still less than half the sequence identity found between GB1 and GA suggesting that there is some relatedness between them indicated by the significant sequence identity.

```

GA      -----MEAVDANSLAQAKEAAIKE--LKQY
GB1     MTYKLILNGKTLKGETTTEAVDAATAEKVFKQY

GA      GIGDYYIKLINNAKTVEGVESLKNEILKALPTE
GB1     A-----NDNGVDGEWTYDDATKTFTVTE

```

Figure 34. MUSCLE sequence alignment of GB1 and GA. Positions colored in red are positions with the exact same amino acid composition. Positions colored in blue are positions with high conservation in amino acid character.

Table 3. List of random small proteins from the protein data bank

PDB Code	Description	Residue Count	% Identity with WT-GB1	% Identity with WT-GA
1JJS	IBiD, Domain of CBP/p300	50	12.2	8.0
1YUF	Type α Transforming Growth Factor	50	8.9	6.0
2L7S	Adrenomedullin	53	13.5	13.0
1ZRP	Rubredoxin	53	14.3	15.2
2JOT	Huwentoxin-XI	55	14.8	15.1
1KMA	Domain-I of the Kazal-type Thrombin Inhibitor Dipetalin	55	12.0	12.2
2MM4	Conserved Golgi Complex-targeting Signal in Coronavirus Envelope Proteins	58	10.7	14.6

```

GA  -MEAVDANSLAQAKEAAIKELKQYGIGDYYI
GB1  -----MTYKLILNGKTLKGETTTEAVDAATA
1JJS  -----GAHMALQDLLRTLKSPSSPQ
1YUF  -----VVSHFNDCPDSHTQFCFHGTCRFLV
2L7S  -----YRQSMNNFQGLRSFGCRFGTCTVQK
1ZRP  AKWVCKICGYIYDEDAGDPDNGISPGTKFEE
2JOT  ----IDTCRLPSDRGRCKASFERWYFNG--R
1KMA  -----FQGNPCECPRALHRVCGSDGNTYSN
2MM4  -----ETGTLIVNSVLLFLAFVVFLVTLAI

GA  KLINNAKTVEGVESLKNEILKALPTE-----
GB1  EKVFKQYANDNGVDGEWTYDDATKTFTVTE--
1JJS  QQQQVLNLIKSNPQLMAAFIKQRTAKYVAN--
1YUF  QEDKPACVCHSGYVGARCEHADLLA-----
2L7S  LAHQIYQFTDKDKDNVAPRSKISPQGYX----
1ZRP  LPDDWVCPICGAPKSEFEKLED-----
2JOT  TCAKFIYGGCGGNGNKFTQEACMKRCAKA--
1KMA  PCMLTCAKHEGNPDLVQVHEGPCDEHDHDF--
2MM4  LTALRLAAYAANIVNVSLVKPTVYVYSRVKNL

```

Figure 35. MUSCLE sequence alignment of GB1 and GA with 7 small random proteins. Sequences where aligned and there were no positions of similarity of identity. Bolded sequences are the target sequences for comparison.

Investigations of both GB1 and GA using PSI-BLAST resulted in a number of common protein sequences primarily in the immunoglobulin-binding cell surface proteins (Table 4). A MUSCLE alignment of the common sequences with both GA and GB1 is found in Figure 36. From the sequence alignment both GB1 and GA align almost fully onto two separate portions of the common sequences. Because the albumin-binding and immunoglobulin G-binding (IgG) domains are found in the protein L or G, it makes sense that both proteins would find their respective relatives. This could indicate that either GA or GB1 may have evolved from a duplication event in either the IgG, albumin-binding

family or other cell surface proteins. The duplication event would then be followed by sequence and functional divergence in which mutations lead to a destabilized structure swapping to a new structure and function. The new function is then stabilized by mutations favorable to the new structure and function. GA and GB1 are similar in that they both bind proteins found in the blood and this new function would be advantageous to the survival of the bacteria. Recently, He *et al.* showed experimentally that these two proteins could be evolved from each other over time and what's even more interesting is that during this synthetic evolution it appears that function is maintained and even at some point towards sequence convergence both functions persisted [373, 390]. In light of He *et al.*'s work it seems plausible that they could have evolved from one another. However, based on the results of Figure 36 which reveals high sequence identity but also that they align on separate portions of the common sequences (Figure 37), it seems unlikely that any of these proteins (Table 4) are ancestral proteins from which GA or GB1 evolved. A true ancestral protein would have overlapping sections of sequence alignment with GA and GB1 and be accompanied with low sequence identity due to evolutionary modification.

Table 4. PSI-BLAST common sequences between GB1 and GA

Accession ID	Code	Description	<u>E-Value</u>		<u>% Identity</u>	
			GB1	GA	GB1	GA
AAA26599.1	AAA4	BBM3XM [<i>Staphylococcus xylosus</i>]	3e-05	5e-10	67	58
AAA26600.1	AAA3	BBXM [<i>Staphylococcus xylosus</i>]	2e-05	1e-10	67	58
AAA26921.1	AAA1	mag [<i>Streptococcus dysgalactiae</i>]	1e-17	2e-05	87	68
AAA86832.1	AAA2	cell surface protein precursor [<i>Streptococcus equi</i> subsp. <i>zooepidemicus</i>]	2e-13	7e-07	73	74
BAD00711.2	BAD1	cell surface protein precursor [<i>Streptococcus equi</i> subsp. <i>zooepidemicus</i>]	7e-14	1e-10	71	68
CAA27638.1	CAA1	protein G [<i>Streptococcus</i> sp.]	1e-16	2e-09	100	58
P06654.1	POA1	Immunoglobulin G-binding protein G; Short=IgG-binding protein G; Flags: Precursor	6e-18	3e-09	100	57
P19909.1	PIA1	Immunoglobulin G-binding protein G; Short=IgG-binding protein G; Flags: Precursor	4e-17	8e-10	100	58
YP_002123072.1	YPA5	IgG binding protein Zag [<i>Streptococcus</i> <i>equi</i> subsp. <i>zooepidemicus</i>]	6e-14	1e-10	71	68
YP_002744821.1	YPA7	Ig, alpha2-macroglobulin and albumin binding protein Zag [<i>Streptococcus</i> <i>equi</i> subsp. <i>zooepidemicus</i>]	9e-14	2e-07	73	74
YP_002746070.1	YPA8	Ig, alpha2-macroglobulin and albumin binding protein Eag [<i>Streptococcus</i> <i>equi</i> subsp. <i>equi</i> 4047]	1e-13	2e-07	73	74
YP_002997067.1	YPA3	Immunoglobulin G-binding protein [<i>Streptococcus dysgalactiae</i> subsp. <i>equisimilis</i>]	5e-18	6e-10	100	58
YP_006013485.1	YPA4	IgG binding protein Zag [<i>Streptococcus</i> <i>dysgalactiae</i> subsp. <i>equisimilis</i>]	2e-17	2e-10	98	58
YP_006042603.1	YPA6	cell surface protein precursor [<i>Streptococcus equi</i> subsp. <i>zooepidemicus</i>]	6e-14	7e-07	73	74
YP_006859906.1	YPA1	Immunoglobulin G-binding protein [<i>Streptococcus dysgalactiae</i> subsp. <i>equisimilis</i>]	4e-18	3e-11	98	58
YP_006904988.1	YPA2	Immunoglobulin G-binding protein G [<i>Streptococcus dysgalactiae</i> subsp. <i>equisimilis</i>]	5e-18	6e-10	100	58
ZP_12571139.1	ZPA1	Immunoglobulin G-binding protein G [<i>Streptococcus dysgalactiae</i> subsp. <i>equisimilis</i>]	1e-17	2e-10	100	58

```

GB1 ------M------
GA1 -----
AAA3 -----MKKKN--IYSIRKLGVGIASVTLGTLISGGVTPAANAAQ----HDEA
AAA4 -----MKKKN--IYSIRKLGVGIASVTLGTLISGGVTPAANAAQ----HDEA
BAD1 -----MEKNKNVSYFLRQSAVGLASVS-AAFLVGTSSVGALDAAT----VLEP
YPA5 MSKFFEKSEGGKMEKNKNVSYFLRQSAVGLASVS-AAFLVGTSSVGALDAAT----VLEP
YPA6 -----MEKNKNVSYFLRQSAVGLASVS-AAFLVGTSSVGALDATT----VLEP
AAA2 -----MEKTKTVSYFLRQSAVGLASVS-AAFLVGTSSVGALDATT----VLEP
YPA8 -----MEKNKKVSYFLRQSAVGLASVS-AAFLVGTSSVGALDAAT----VLEP
YPA7 -----MEKNKKVSYFLRQSAVGLASVS-AAFLVGTSSVGALDAAT----VLEP
AAA1 -----MEKEKKVKYFLRKSAFGLASVS-AAFLVGTAVVNAEESTVSPVTVATD
POA1 -----MEKEKKVKYFLRKSAFGLASVS-AAFLVGSTVF-AVDSPI----EDTP
YPA1 MCWHIKIKKGEKMEKEKKVKYFLRKSAFGLASVS-AAFLVGSTVF-AVDSPI----EDTP
YPA4 -----MEKEKKVKYFLRKSAFGLASVS-AAFLVGSTVF-AVDSPI----EDTP
YPA2 -----MEKEKKVKYFLRKSAFGLASVS-AAFLVGSTVF-AVDSPI----EDTP
YPA3 -----MEKEKKVKYFLRKSAFGLASVS-AAFLVGSTVF-AVDSPI----EDTP
CAA1 -----EFNKYGVSDDYK-----
P1A1 -----MEKEKKVKYFLRKSAFGLASVS-AAFLVGSTVF-AVDSPI----EDTP

GB1 -----
GA1 -----
AAA3 VDANFDQFNKYGVSDYYKNLINNAKTVEGVKDLQAQVVESAKKARISEATDGLSDFLKSQ
AAA4 VDANFDQFNKYGVSDYYKNLINNAKTVEGVKDLQAQVVESAKKARISEATDGLSDFLKSQ
BAD1 TTAF-----IREAVREINQ-----LSDDYADNQE-LQAVLANAGVEALAADTVDQ
YPA5 TTAF-----IREAVREINQ-----LSDDYADNQE-LQAVLANAGVEALAADTVDQ
YPA6 TTAF-----IREAVREINQ-----LSDDYADNQE-LQAVLANAGVEALAADTVDQ
AAA2 TTAF-----IREAVREINQ-----LSDDYADNQE-LQAVLANAGVEALAADTVDQ
YPA8 TTAF-----IREAVREINQ-----LSDDYADNQE-LQAVLANAGVEALAADTVDQ
YPA7 TTAF-----IREAVREINQ-----LSDDYADNQE-LQAVLANAGVEALAADTVDQ
AAA1 AVTT-----SKEALAIINK-----LSEDNLNNLD-IQEVLAQAGRDILASDSADT
POA1 IIRN-----GGELTNLLGNSETTLALRNEESATAD-LTAAAVADTVAAAAAENAGA
YPA1 IIRN-----GGELTNLLGNSETTLALRNEESATAD-LTAAAVADTVAAAAAENAGA
YPA4 IIRN-----GGELTNLLGNSETTLALRNEESATAD-LTAAAVADTVAAAAAENAGA
YPA2 IIRN-----GGELTNLLGNSETTLALRNEESATAD-LTAAAVADTVAAAAAENAGA
YPA3 IIRN-----GGELTNLLGNSETTLALRNEESATAD-LTAAAVADTVAAAAAENAGA
CAA1 -----NLINNAKTVEGVKDLQAQVVESAKKARISEATDGLSDFLKSQ
P1A1 IIRN-----GGELTNLLGNSETTLALRNEESATAD-LTAAAVADTVAAAAAENAGA

```

Figure 36. MUSCLE alignment of common PSI-BLAST protein sequences. Sequence colored in red is the aligned portion for GA (3-helical bundle) and the sequence in blue is the aligned portion for GB1 (4 β + α fold). Bolded sequences are the target sequences for comparison.

```

GB1 -----
GA1 -----
AAA3 TPAEDTVKSIELAEAKVLANRELDKYGVSDYHKNL-----
AAA4 TPAEDTVKSIELAEAKVLANRELDKYGVSDYHKNL-----
BAD1 AKAALDKAKAAVAGVQLDEARR-----EAYRAI-----
YPA5 AKAALDKAKAAVAGVQLDEARR-----EAYRAI-----
YPA6 AKAALDKAKAAVAGVQLDEARR-----EAYRTI-----
AAA2 AKAALDKAKAAVAGVQLDEARR-----EAYRTI-----
YPA8 AKAALDKAKAAVAGVQLDEARR-----EAYRTI-----
YPA7 AKAALDKAKAAVAGVQLDEARR-----EAYRTI-----
AAA1 IKALLAEVTAEVTRLNEEKMAR-----DAVDKAIADAAAFSELKDAQLKAYED-----
POA1 AAWEAAAAADALAKAKADALKEFNKYGVSDYYKNL-----
YPA1 AAWEAAAAADALAKAKADALKEFNKYGVSDYYKNL-----
YPA4 AAWEAAAAADALAKAKADALKEFNKYGVSDYYKNL-----
YPA2 AAWEAAAAADALAKAKADALKEFNKYGVSDYYKNLINNAKTVEGVKDLQAQVVESAKKAR
YPA3 AAWEAAAAADALAKAKADALKEFNKYGVSDYYKNLINNAKTVEGVKDLQAQVVESAKKAR
CAA1 TPAEDTVKSIELAEAKVLANRELDKYGVSDYHKNL-----
P1A1 AAWEAAAAADALAKAKADALKEFNKYGVSDYYKNLINNAKTVEGVKDLQAQVVESAKKAR

GB1 -----
GA1 -----
AAA3 -----INNAKTVEGV
AAA4 -----INNAKTVEGV
BAD1 -----NALSDQHESDQKV
YPA5 -----NALSDQHESDQKV
YPA6 -----NALSDQHESDQKV
AAA2 -----NALSDQHESDQKV
YPA8 -----NALSDQHKSDQKV
YPA7 -----NALSDQHESDQKV
AAA1 -----LAKLAADTDL
POA1 -----INNAKTVEGI
YPA1 -----INNAKTVEGV
YPA4 -----INNAKTVEGV
YPA2 ISEATDGLSDFLKSQTPAEDTVKSIELAEAKVLANRELDKYGVSDYHKNLINNAKTVEGV
YPA3 ISEATDGLSDFLKSQTPAEDTVKSIELAEAKVLANRELDKYGVSDYHKNLINNAKTVEGV
CAA1 -----INNAKTVEGV
P1A1 ISEATDGLSDFLKSQTPAEDTVKSIELAEAKVLANRELDKYGVSDYHKNLINNAKTVEGV

```

Figure 36. Continued.

```

GB1 -----
GA1 -----MEAVDANSIAQAKEAAIKEL
AAA3 K-DLQAQVVESAKKARISEATDGLSDFLKS-----QTPAEITVKSIEIABAKYLAANKEL
AAA4 K-DLQAQVVESAKKARISEATDGLSDFLKS-----QTPAEITVKSIEIABAKYLAANKEL
BAD1 QLALVAAAAKVADAVSVDQVNAAINDVREE-----IAGITVLAARALIKAKPAATINEL
YPA5 QLALVAAAAKVADAVSVDQVNAAINDVREE-----IAGITVLAARALIKAKPAATINEL
YPA6 QLALVAAAAKVADAASVDQVNAAIND-----AHTAIAALITVAALIPAKPAATINEL
AAA2 QLALVAAAAKVADAASVDQVNAAIND-----AHTAIAALITVAALIPAKPAATINEL
YPA8 QLALVAAAAKVADAASVDQVNAAIND-----AHTAIAALITVAALIPAKPAATINEL
YPA7 QLALVAAAAKVADAASVDQVNAAIND-----AHTAIAALITVAALIPAKPAATINEL
AAA1 DLDVAKIINDYTTKVENAKTAEDVKKIFEESQNEVTRIKTAKAIYAAALIPAKPAATINEL
POA1 K-DLQAQVVESAKKARISEATDGLSDFLKS-----QTPAEITVKSIEIABAKYLAANKEL
YPA1 K-DLQAQVVESAKKARISEATDGLSDFLKS-----QTPAEITVKSIEIABAKYLAANKEL
YPA4 K-DLQAQVVESAKKARISEATDGLSDFLKS-----QTPAEITVKSIEIABAKYLAANKEL
YPA2 K-DLQAQVVESAKKARISEATDGLSDFLKS-----QTPAEITVKSIEIABAKYLAANKEL
YPA3 K-DLQAQVVESAKKARISEATDGLSDFLKS-----QTPAEITVKSIEIABAKYLAANKEL
CAA1 K-DLQAQVVESAKKARISEATDGLSDFLKS-----QTPAEITVKSIEIABAKYLAANKEL
P1A1 K-DLQAQVVESAKKARISEATDGLSDFLKS-----QTPAEITVKSIEIABAKYLAANKEL

GB1 -----
GA1 KQYGIGDYIYIKLINNAKTVEGVESLKNEILKALPTE-----
AAA3 LKQYALGYYKILINNAKTVEGVKALILGILAAALIKTDT-----
AAA4 LKQYALGYYKILINNAKTVEGVKALILGILAAALIKTDTYKLIL-----
BAD1 KQYALGYYKILINNAKTVEGVKALILGILAAALIKTDTVEV-----
YPA5 KQYALGYYKILINNAKTVEGVKALILGILAAALIKTDTVEV-----
YPA6 KQYALGYYKILINNAKTVEGVKALILGILAAALIKTDTVEV-----
AAA2 KQYALGYYKILINNAKTVEGVKALILGILAAALIKTDTVEV-----
YPA8 KQYALGYYKILINNAKTVEGVKALILGILAAALIKTDTVEV-----
YPA7 KQYALGYYKILINNAKTVEGVKALILGILAAALIKTDTVEV-----
AAA1 KQYALGYYKILINNAKTVEGVKALILGILAAALIKTDT-----
POA1 LKQYALGYYKILINNAKTVEGVKALILGILAAALIKTDT-----
YPA1 LKQYALGYYKILINNAKTVEGVKALILGILAAALIKTDT-----
YPA4 LKQYALGYYKILINNAKTVEGVKALILGILAAALIKTDT-----
YPA2 LKQYALGYYKILINNAKTVEGVKALILGILAAALIKTDT-----
YPA3 LKQYALGYYKILINNAKTVEGVKALILGILAAALIKTDT-----
CAA1 LKQYALGYYKILINNAKTVEGVKALILGILAAALIKTDTYKLILNGKTLKGETTTEAVDAA
P1A1 LKQYALGYYKILINNAKTVEGVKALILGILAAALIKTDTYKLILNGKTLKGETTTEAVDAA

```

Figure 36. Continued.


```

GB1 -----
GA1 -----
AAA3 -----YKLILNGKTLKG
AAA4 -----NGKTLKG
BAD1 -----IDAAELTPALTSYKLVIKGATFSG
YPA5 -----IDAAELTPALTSYKLVIKGATFSG
YPA6 -----IDAAELTPALTSYKLVIKGATFSG
AAA2 -----IDAAELTPALTSYKLVIKGATFSG
YPA8 -----IDAAELTPALTSYKLVIKGATFSG
YPA7 -----IDAAELTPALTSYKLVIKGATFSG
AAA1 -----
POA1 -----YKLILNGKTLKG
YPA1 -----YKLILNGKTLKG
YPA4 -----YKLILNGKTLKG
YPA2 -----YKLILNGKTLKG
YPA3 -----YKLILNGKTLKG
CAA1 TAEKVFKQYANDNGVDGEWYDDATKTFTVTEKPEVIDASELTPAVTTYKLVINGKTLKG
P1A1 TAEKVFKQYANDNGVDGEWYDDATKTFTVTEKPEVIDASELTPAVTTYKLVINGKTLKG

GB1 -----
GA1 -----
AAA3 ETTTEAVDAATARSFNFPILENSSSVPGD-----PLESTCRHA-----
AAA4 ETTTEAVDAATARSFNFPILENSSSVPGDPLESTCMHVEHDAEENVEHDAEENVEHDAEE
BAD1 ETATKAVDAAAVAEQT-FRDYANKNGVDGV-----WAYDAATKTF-----
YPA5 ETATKAVDAAAVAEQT-FRDYANKNGVDGV-----WAYDAATKTF-----
YPA6 ETATKAVDAAAVAEQT-FRDYANKNGVDGV-----WAYDAATKTF-----
AAA2 ETATKAVDAAAVAEQT-FRDYANKNGVDGV-----WAYDAATKTF-----
YPA8 ETATKAVDAAAVAEQT-FRDYANKNGVDGV-----WAYDAATKTF-----
YPA7 ETATKAVDAAAVAEQT-FRDYANKNGVDGV-----WAYDAATKTF-----
AAA1 -----
POA1 ETTTEAVDAATAAEKV-FKQYANDNGVDGE-----WTYDDATKTF-----
YPA1 QTTTEAVDAATAAEKV-FKQYANDNGVDGE-----WTYDDATKTF-----
YPA4 QTTTEAVDAATAAEKV-FKQYANDNGVDGE-----WTYDDATKTF-----
YPA2 ETTTEAVDAATAAEKV-FKQYANDNGVDGE-----WTYDDATKTF-----
YPA3 ETTTEAVDAATAAEKV-FKQYANDNGVDGE-----WTYDDATKTF-----
CAA1 ETTTEAVDAATAAEKV-FKQYANDNGVDGE-----WTYDDATKTF-----
P1A1 ETTTEAVDAATAAEKV-FKQYANDNGVDGE-----WTYDDATKTF-----

```

Figure 36. Continued.

```

GB1 -----
GA1 -----
AAA3 -----S
AAA4 NVEHDAEENVEHDAEENVEENVEENVEENVEENVEENVEENVEENVEENVEENVSS
BAD1 -----T
YPA5 -----T
YPA6 -----T
AAA2 -----T
YPA8 -----T
YPA7 -----T
AAA1 -----
POA1 -----T
YPA1 -----T
YPA4 -----T
YPA2 -----T
YPA3 -----T
CAA1 -----T
P1A1 -----T

GB1 -----TYKLIILNGKTLKGETTTEAVDAATAEKVFKQYAND
GA1 -----
AAA3 FAQAPKEEDNNKPGKEDNNKPGKEDNNKPGKEDNNKPGKEDNNKPGKEDNNKPGKEDGNNK
AAA4 FAQAPKEEDNNKPGKEDNNKPGKEDNNKPGKEDNNKPGKEDNNKPGKEDNNKPGKEDGNNK
BAD1 VTEQPVAET-----IEAAELTPALTTYRLVIKGVTFSGETATKAVDAATAEQTFRQYAND
YPA5 VTEQPVAET-----IEAAELTPALTTYRLVIKGVTFSGETATKAVDAATAEQTFRQYAND
YPA6 VTEQPVAET-----IEAAELTPALTTYRLVIKGVTFSGETATKAVDAATAEQAFRQYAND
AAA2 VTEQPVAET-----IEAAELTPALTTYRLVIKGVTFSGETATKAVDAATAEQAFRQYAND
YPA8 VTEQPVAET-----IEAAELTPALTTYRLVIKGVTFSGETATKAVDAATAEQTFRQYAND
YPA7 VTEQPVAET-----IEAAELTPALTTYRLVIKGVTFSGETSTKAVDAATAEQTFRQYAND
AAA1 -----IDAPELTPALTTYKLVINGKTLKGETTTKAVDAETA EKAFKQYANE
POA1 VTEKP--EV-----IDASELTPAVTTYKLVINGKTLKGETTTKAVDAETA EKAFKQYAND
YPA1 VTEKP--EV-----IDAPELIPAVTTYKLVINGKTLKGETTTKAVDAETA EKAFKQYAND
YPA4 VTEKP--EV-----IDASELTPAVTTYKLVINGKTLKGETTTKAVDAETA EKAFKQYAND
YPA2 VTEKP--EV-----IDASELTPAVTTYKLVINGKTLKGETTTKAVDAETA EKAFKQYAND
YPA3 VTEKP--EV-----IDASELTPAVTTYKLVINGKTLKGETTTKAVDAETA EKAFKQYAND
CAA1 VTEKP--EV-----IDASELTPAVTTYKLVINGKTLKGETTTKAVDAETA EKAFKQYAND
P1A1 VTEKP--EV-----IDASELTPAVTTYKLVINGKTLKGETTTKAVDAETA EKAFKQYAND

```

Figure 36. Continued.

Figure 36. Continued.

```

GB1 -----
GA1 -----
AAA3 GTTVFGGLSLALGAALLAGRRREL
AAA4 GTTVFGGLSLALGAALLAGRRREL
BAD1 TAAALAIMA-GAGALAVTSKRQQD
YPA5 TAAALAIMA-GAGALAVTSKRQQD
YPA6 TAAALAIMA-GAGALAVTSKRQQD
AAA2 TAAALAIMA-GAGALAVTSKRQQD
YPA8 TAAALAIMA-SAGALAVTSKRQQD
YPA7 TAAALAIMA-SAGALAVTSKRQQD
AAA1 TAAALAVMA-GAGALAVASKRKED
POA1 TAAALAVMA-GAGALAVASKRKED
YPA1 TAAALAVMA-GAGALAVASKRKED
YPA4 TAAALAVMA-GAGALAVASKRKED
YPA2 TAAALAVMA-GAGALAVASKRKED
YPA3 TAAALAVMA-GAGALAVASKRKED
CAA1 TAAALAVMA-GAGALAVASKRKED
P1A1 TAAALAVMA-GAGALAVASKRKED

```

Figure 36. Continued.

Figure 37. Condensed and truncated schematic of MUSCLE alignment of common PSI-BLAST protein sequences. Sequences colored in red are aligned portions with GA (3-helical bundle) and the sequences in blue are aligned portions with GB1 (4 β + α fold). The letters are not meant to be visible. The purpose of the figure is to show the overlapping sections using color.

Due to their low sequence similarity it is more likely that GB1 and GA evolved from a common ancestral type protein. To explore where this divergence may have begun further iterative searches were done to look at other distantly related proteins further back

in evolution. Sequence identity and PSI-BLAST E-values less than $5e-04$ are a good general method for identifying related sequences. Higher sequence identity indicates more relatedness and closer in evolution. Because the first set of PSI-BLAST common sequences had high sequence identity without converging onto a common sequence alignment, there could be other more distantly related ancestral sequences and functions from which they may have diverged. A protein that is still related but further back on the evolutionary pathway will have a lower sequence identity due to mutational variance indicating an indirect relationship but still linked via degrees of separation.

To explore other families a phylogenetic tree was constructed using the initial PSI-BLAST results of GA (Figure 38) and GB1 (Figure 39) independently. The initial sequences of both GA and GB1 was linked to a gram-positive signaling peptide protein of the YSIRK family. The YSIRK family of proteins consists of any proteins that contain the YSIRK cell surface targeting motif. This sequence allows for membrane bound, peptidoglycan anchored or transmembrane proteins to be directed and secreted to the cell surface [391]. Although both GB1 and GA were found to be related to a YSIRK type protein a further BLAST search using these proteins did not result in picking up either the GB1 related YSIRK or the GA related YSIRK proteins. We believe that this is most-likely due to the classification of YSIRK family proteins based on the presence of the motif and not necessarily by function or structural components. There are a significant number of proteins that function on the cell surface and this indicates that the two proteins may contain the signaling motif but not be directly related in evolution. It is interesting to note that from the phylogenetic analysis of GB1, the proteins genetically evolved to 88, 95 and 98 % identity of each other showed to be more closely related to

the YSIRK protein. This indicates that the synthetically converged sequence between GB1 and GA may be a close approximation for an evolutionary intermediate between the YSIRK protein and GB1. However, this is only speculative and those same sequences appear to be more closely related to GA in the phylogenetic analysis of GA indicating that the synthetic mutation may have favored the GA sequence over the GB1 sequence. From the phylogenetic trees of GB1 and GA proteins the YSIRK proteins were selected because they are more distant yet still related. We also believe the YSIRK signaling motif may be an inherited motif from the ancestral protein. The selected YSIRK proteins were used as the target sequence for further PSI-BLAST searches to see if there is a convergence between the two sequences.

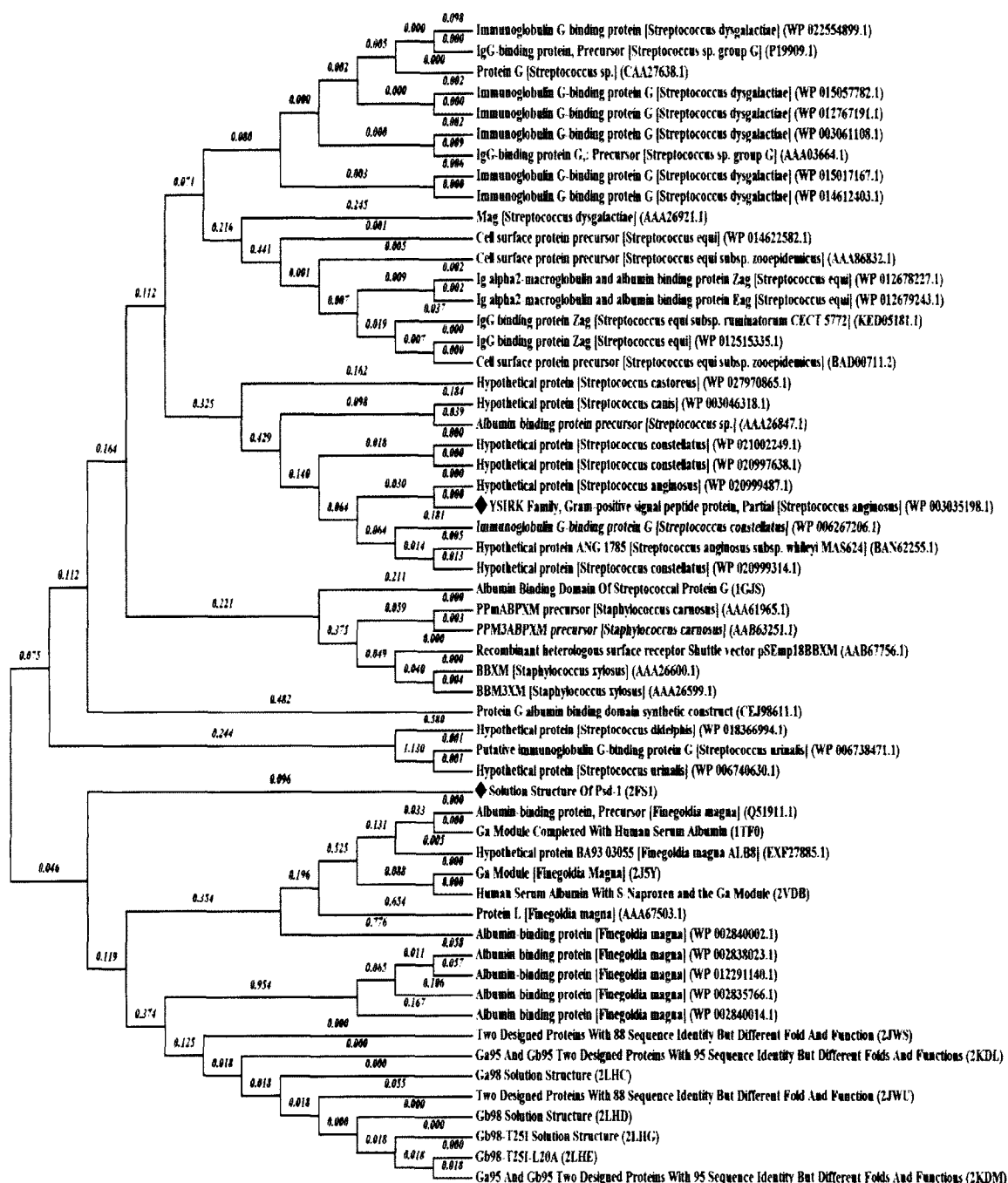


Figure 38. Phylogenetic tree of related sequences to GA from PSI-BLAST. Positions indicated by the diamonds are GA (red) and YSIRK family gram-positive signal peptide protein (blue). Evolutionary distances are shown above each link. The phylogenetic tree was constructed using the program, Molecular Evolutionary Genetics Analysis (MEGA Ver. 6).

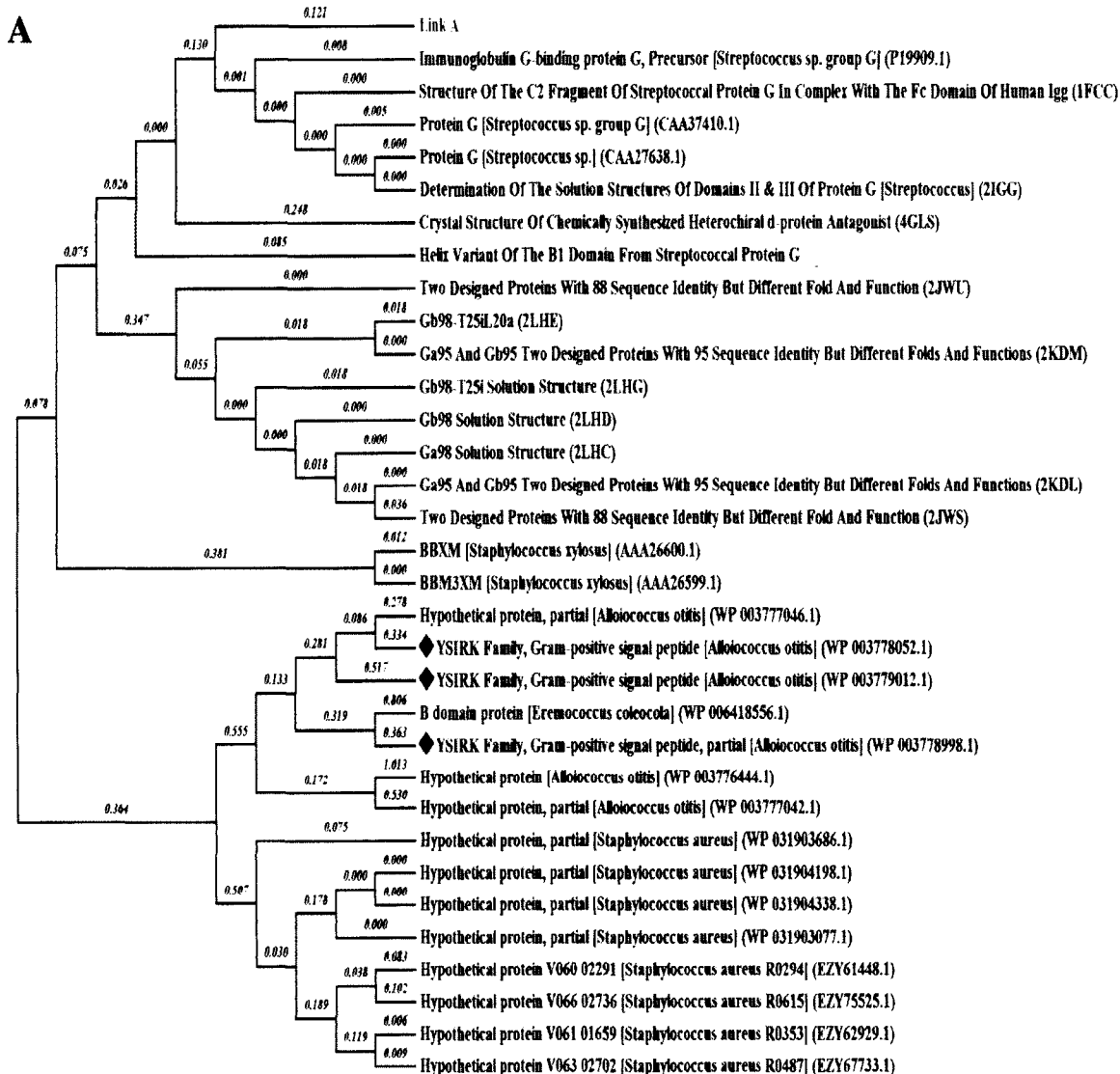


Figure 39. Phylogenetic tree of related sequences to GB1 from PSI-BLAST. Positions indicated by the diamonds are GB1 (red) and YSIRK family gram-positive signal peptide proteins (blue). Panel B is connected to panel A at the Link A position. Evolutionary distances are shown above each link. The phylogenetic tree was constructed using the program, Molecular Evolutionary Genetics Analysis (MEGA Ver. 6).

B

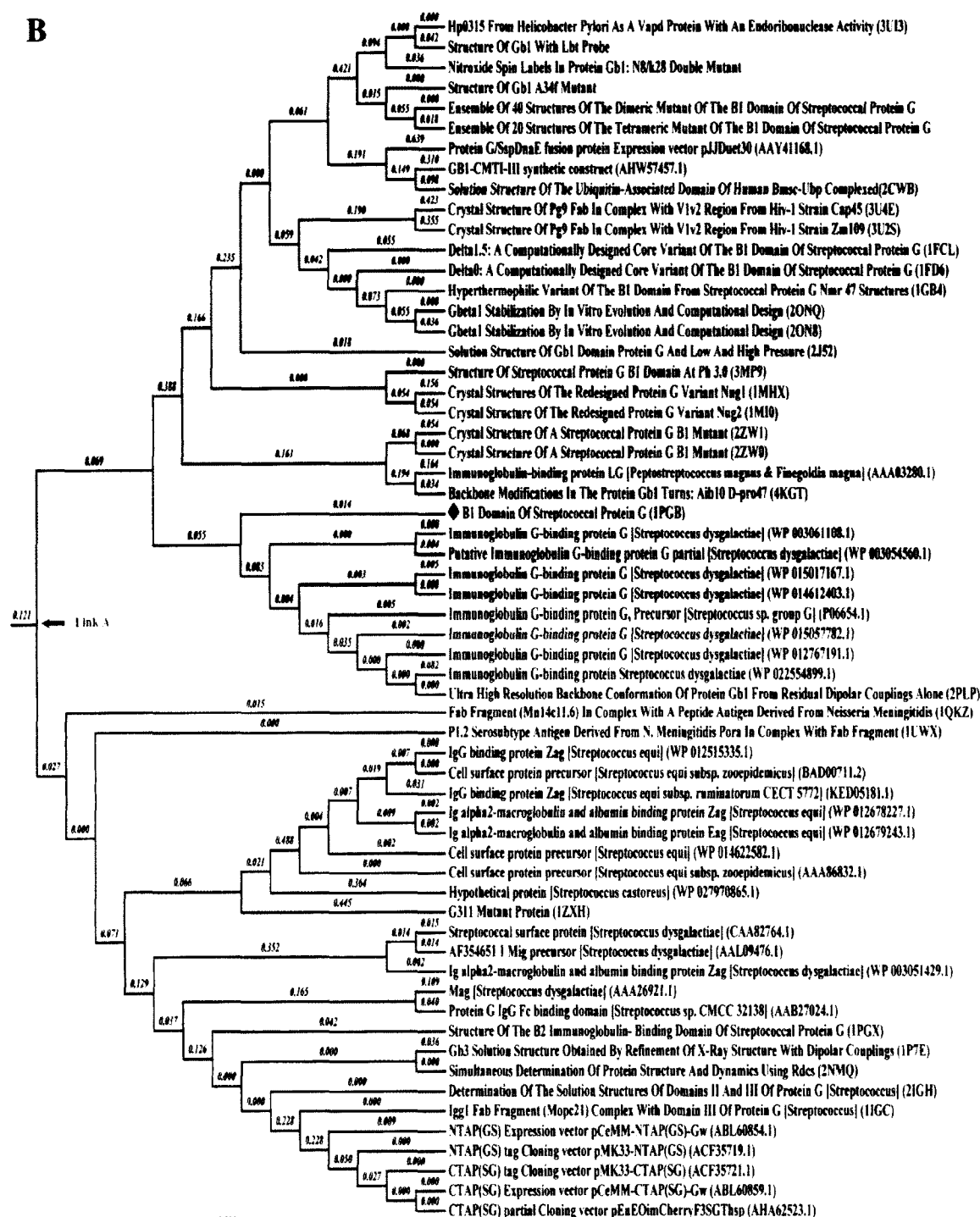


Figure 39. Continued.

Using the YSIRK proteins as our query sequence resulted in a much more diverse group of related proteins with variable functions. The results of the PSI-BLAST search resulted in a convergence on 12 common mucus-binding proteins. From the 12 sequences a phylogenetic tree was constructed to determine their relationship to each other in evolution (Figure 40). Further analysis using MUSCLE indicated that the two YSIRK family proteins were more closely related to the mucus-binding protein sequence indicated with a red diamond in Figure 40. They are more distantly related to the oldest sequence of the 12 proteins (accession code: YP_005861639.1). A MUSCLE alignment of the YSIRK family proteins with the 12 mucus-binding proteins showed a convergence of sequence alignment with portions of both sequences aligned on similar regions of the mucus-binding proteins (Figure A1). This indicates that GA and GB1's evolutionary pathways may have diverged from a common ancestral mucus-binding protein (WP_000287308.1) and further evolved independently. This convergence of sequence and the distance in evolutionary relatedness could account for the significant difference in sequence identity between GB1 and GA. To see how they related to each other a MUSCLE sequence alignment of GB1 and GA with the 12 mucus-binding proteins was constructed (Figure A2). From the sequence alignment of GB1 and GA with the 12 mucus-binding proteins it is clear that these proteins are very distantly related as there are small regions of overlap scattered throughout the mucus-binding protein which is what is seen from a distantly related evolutionary ancestor.

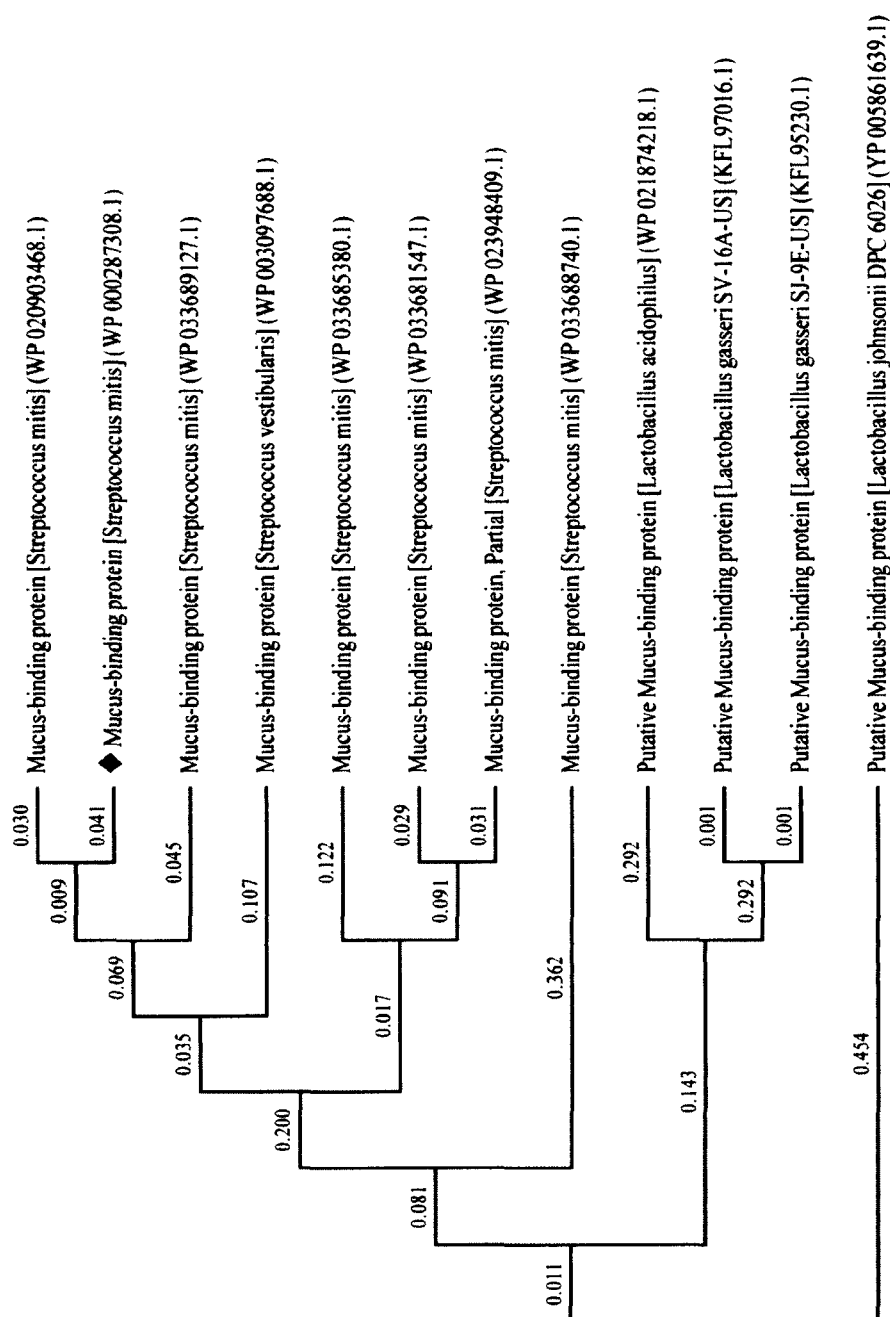


Figure 40. Phylogenetic tree of 12 mucus-binding proteins. Position indicated by the red diamond is the sequence of highest identity between the YSIRK family gram-positive signal peptide proteins of GA and GB1. Evolutionary distances are shown above each link. The phylogenetic tree was constructed using the program, Molecular Evolutionary Genetics Analysis (MEGA Ver. 6).

Both GB1 and GA are truncated domains of the immunoglobulin G-binding protein G and the albumin-binding protein, respectively. We wanted to see if there was a relationship between the full-length immunoglobulin G-binding protein G and the albumin-binding protein sequences found in BLAST and the YSIRK family proteins found with just the individual domains. Both full length proteins were shown to be related to their respective YSIRK family protein after a PSI-BLAST analysis. However, there was a reduction in sequence identity in comparison with the individual domains indicating a more distant relationship. A phylogenetic analysis of all the proteins further supported the evolutionary relationship between each of the related proteins (Figure 41). From the phylogeny tree we see that the GA module is closely related to the albumin-binding protein. There was a speciation event that leads to separation from the YSIRK family protein related to GA followed by divergence from the mucus-binding protein. Similarly, GB1 is closely related to the IgG-binding protein G. They both resulted from a speciation event leading to separation of the YSIRK family and GB1 related proteins. The tree clearly shows that both GA and GB1 arose from a common ancestral protein which we postulate to be an older mucus-binding protein. In this particular analysis it seems that the selected ancestral protein is more closely related to the GA proteins rather than the GB1. This supports our previous sequence identity findings between the YSIRK family and the mucus-binding proteins. A schematic of the overall computational design was constructed and the corresponding results to show how the ancestor was determined (Figure 42). In the schematic we can see that the percent identity is greater with a significance E-value less than 0.005 indicating more relatedness.

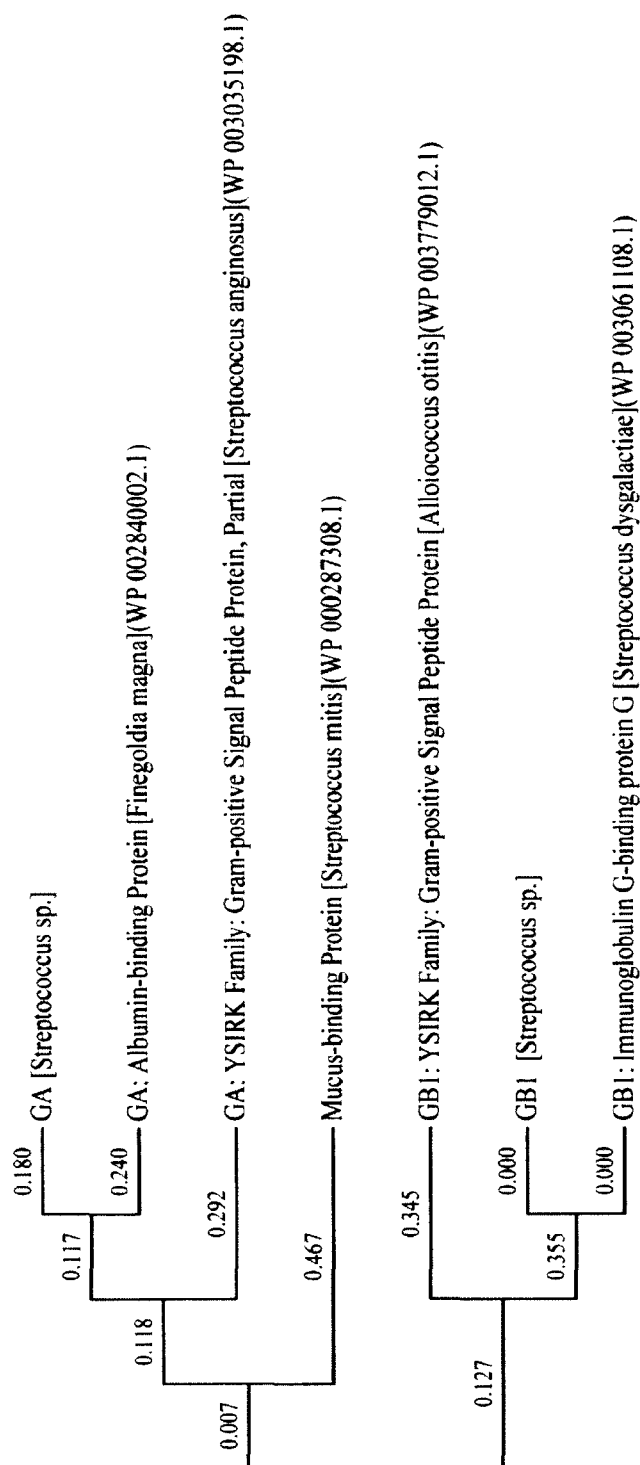


Figure 41. Phylogenetic analysis of all the related proteins from GB1 and GA.

Evolutionary distances are shown above each link. The phylogenetic tree was constructed using the program, Molecular Evolutionary Genetics Analysis (MEGA ver. 6).

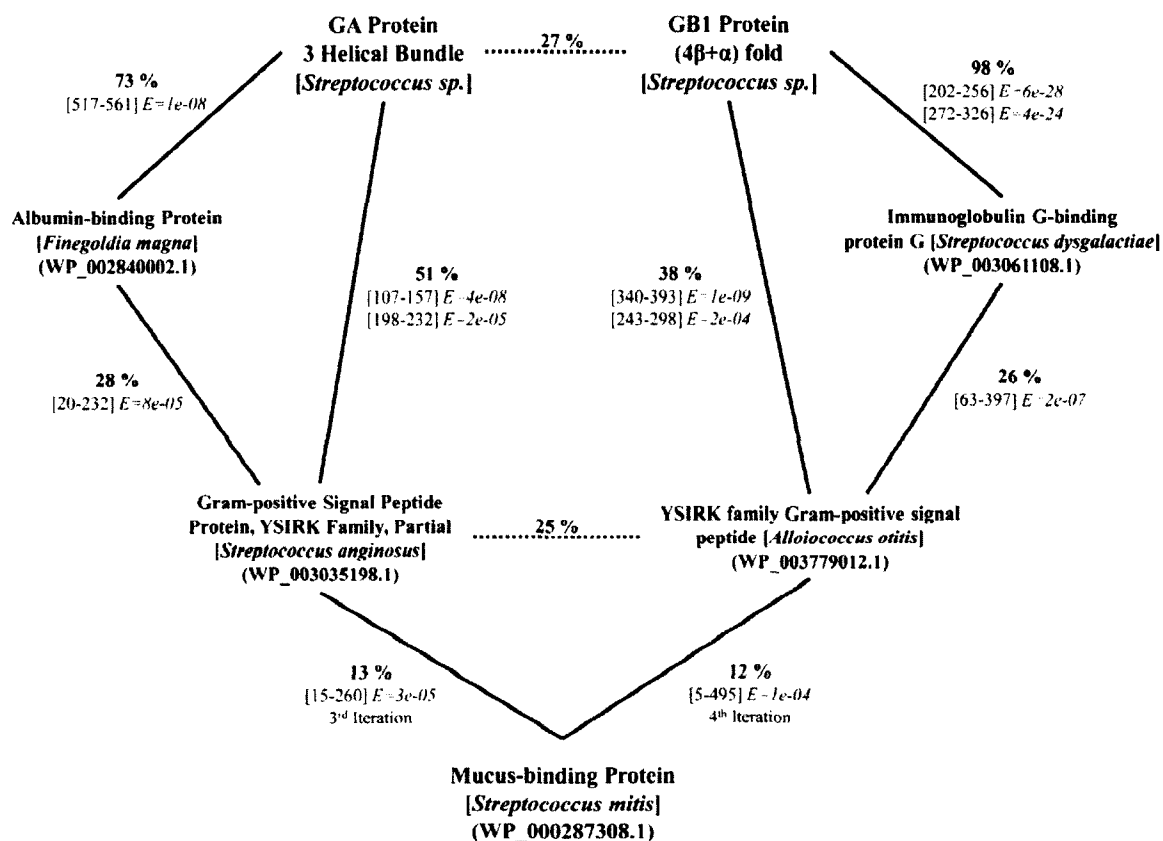


Figure 42. Schematic of the computational analysis and corresponding results. Solid lines indicate sequences related via PSI-BLAST analysis. Each solid line is labeled with the percent identity, regions of alignment coverage and e-values. In some cases PSI-BLAST was able to align the sequence multiple times with e-value less than $5e-03$ resulting in multiple significance regions. Dotted lines represent corresponding percent identity made using MUSCLE alignments and do not indicate direct evolutionary relatedness. In the last two searches multiple iterations were required in order to converge on the mucus-binding proteins and the number of iterations is denoted.

To further investigate positions that may be conserved in evolution between both GB1 and GA a MUSCLE alignment of all the related proteins was constructed (Figure 43). As was expected we find that there are portions of both GB1 and GA that overlap with the ancestral mucus-binding protein which further supports that they diverged from this family of proteins. Surprisingly, there is a single completely conserved Thr11 (GB1) and Thr38 (GA) through the evolutionary process which could mean that this position could be fundamentally critical in the function or structural stability of both proteins. An experimental investigation into this particular residue to determine its role would prove quite valuable in understanding its importance in evolutionary conservation. There are 12 residues of the original GA and GB1 proteins that aligned completely with all related sequences. Interestingly, if we look at portions aligned to only the GA or GB branches we find large portions of overlap and, more importantly, we find a significant number of evolutionarily conserved positions on both sides. These conserved positions may play critical roles in stabilizing the fold or function functions. From the condensed sequences of GB1 and GA we can see the distribution of aligned positions throughout the sequence (Figure 44). There are a surprising number of evolutionarily conserved positions scattered throughout both proteins. It is interesting that the completely aligned regions in GA (residues 32-40) and GB1 (residues 1-13) fall on different portions of their sequence. For GA we find it mainly centralized where as in GB1 it is found at the start of the sequence. These particular sequences appear to have very little variation with an almost completely conserved LINxxKTV sequence throughout evolution. It seems that nature has cleverly reused the same sequence in the evolution of two separate proteins by simple

translocation. This finding seems to indicate that, when possible, this method of reusing sequences is more efficient than mutating every position to achieve the required changes.

```

      GA  -----
      GB1 -----
      GA-AB -----MKINKKLLMAALAGAIVVSGGVSTYAEGETATTPAKKSQPAATLLTPEDA
      GB1-IGG -----
      GA-YSIRK -----
      GB1-YSIRK -----
      AMBP MYSRMEKYHGRRARFQSIRKYSFGAASVLLGTALLLGANAVKADETSTASTKTSEVTNSD

      GA  -----
      GB1 -----
      GA-AB RAKAEKEAADKKAKKEAAKRLSESKITAKNVIDGLDLSKFQKRI FKNKVEDATSRKQVK
      GB1-IGG -----MGSTVFAVDSPIEDT
      GA-YSIRK -----
      GB1-YSIRK -----MVGKNNYKTR
      AMBP KQKPDSAITTPVVEELPELKIDAVKADEKPEVKEEAKPVAEKEVTDKAATEKSDKEQADK

      GA  -----
      GB1 -----
      GA-AB QLVNEARELVDYISNAKMSLGLKEGQSFEELISKKFDQATTKEQVDSAKQFAAEPEGYKK
      GB1-IGG PIIRNGGELTNLLGNSETTLALRNEESATADLTAAAVA-----
      GA-YSIRK ---MENKKMKYYLRKSAFGLAAVSASVLVGVTTVSAQVTTTRAQAAKLREEATQKISELEK
      GB1-YSIRK TEKAANKKQRFSLKKLSVGVASVAVGTTFLSNTDAVS-----
      AMBP KEVAKEKTDKESPKKAATEKAQDEVKTVLTQLTSEAEVMASVASNFSDKEVKDEAAKKEL

```

Figure 43. MUSCLE alignment of all related proteins. Sequence Alignment of GA and GB1 with IgG-binding Protein G (GB1-IGG), albumin-binding (GA-AB), GA gram-positive signal peptide protein (GA-YSIRK), GB1 gram-positive signal peptide protein (GB1-YSIRK) and the possible common ancestral mucus-binding protein (AMBP). Residues common to all sequences are indicated in red. Residues common to only the GA branch are indicated in green and those common to the GB1 branch alone are indicated in purple. Highlighted in yellow are positions that are similar in identity or character.


```

GA -----
GB1 -----
GA-AB DQELKAAKEAAKEEIKKLDLSDAQKNNFNSKVDEATDVDGVNKKVKEEAKDFLNYLIESKG
GB1-IGG -----
GA-YSIRK TIDNKT-----
GB1-YSIRK -----
AMBP AVTIEAVKLEAAKSNDLLSSDASKDQMVAVQNRLSAAIEAVYTEMKRAGHAGKVESVLAA

GA -----
GB1 -----
GA-AB KIKASDNEDMVKKTAKFESALTKADLDNVVTEINKEIEEGKKKTDPKEEEEAKKLAAAK
GB1-IGG -----
GA-YSIRK -----
GB1-YSIRK -----
AMBP TASKITGKDVLDGETVNAVTVNAYVDMNADNTKPVGWGFDTTISTSTLKAGSITKIELTN

GA -----
GB1 -----
GA-AB EAAKEEINKLDLSDAQKNNFNSNVDKATDVDGVNKKVKEEAKDFLNYLIESKGKIKASDDE
GB1-IGG -----
GA-YSIRK -----
GB1-YSIRK -----
AMBP LAELGGGLAVNTEIRETDGTVVGVKKSIDYKTTTGNNNNKSTPYWGQRTQRGMTYDQ RVA

GA -----
GB1 -----
GA-AB DMVKKTAKFDAAALTKADLDKVVAEINKEIEAQEKEKQKAKELAEAKENAKKHIDELKHL
GB1-IGG -----
GA-YSIRK -----
GB1-YSIRK -----
AMBP EQPAVANETGTYTYNIEWNDKVVDYPNVSFSGASNLGSGYYAPEISKDTPYTATIKIDGR

GA -----
GB1 -----
GA-AB SDKAKELAKKDIDEATTIDEIKDIVAKADVMRKTAE-----KEEAELKLAALKDA--KE
GB1-IGG -----DTVAAAAAEN-----AGAAWEAAAAADA--LA
GA-YSIRK ---QLDSVKKEIQNAVDRDKIQELSNQADQIVSAQAE-----KEALIKSEKKLADAWELE
GB1-YSIRK -----AEENP-----DRNIEQLQALETE--KE
AMBP TVLEHTYTRKGQQPNYQKQTSASLSENNGLTYLNNEQTGRSDSIVLKTDSDVRYGVGSK

GA -----
GB1 -----
GA-AB KAIEAIRK---EGVKSPLYEDLINKAKTIDGVNALRDQII EAHKASNPG-----
GB1-IGG KAKADALKEFNKYGVSDYYKNLINNAKTVEGVKDLQAQVVESAKKARIS-----
GA-YSIRK AAKDAALKELNQYGVSDYYKKLVNSAKTVAGVKKLQAQVVESAKKARVS-----
GB1-YSIRK FAIAEVQA---AGFTDQKYVDWINAQETVQDVNGVKREILTTPDEEQSGIEEETSV---
AMBP FTIKLPNADFTEFKELEGSSNFVNGLNTASTINPNKGDSITYRPASRWANVKANENNVWI

```

Figure 43. Continued.

```

GA -----MEAVDANSLAQAKEAAI-----
GB1 -----
GA-AB -----ITIDEWLLKNAKEDAI-----
GB1-IGG -----EATDGLSDFLKSQTPAEDTVKSIELAEAKVLAN-----
GA-YSIRK -----EATDGLSGFLKSQTPAEDTIKSLELSEAKTLAL-----
GB1-YSIRK -----DSESDFNAPDFDWSGNDEAVAAEAEELQAAKESAI-----
AMBP LNDGRDSGFTLTPLRLISPTLELELTVTEGAIQEGSVVSMPLQSLGIEKVIKDKTLTSEYSK

GA -----
GB1 -----
GA-AB -----
GB1-IGG -----
GA-YSIRK -----
GB1-YSIRK -----
AMBP ITYENGLIKEGYVGNDKTAATLTVSGGESVNGEKEDVATTVPNGWSVKGDGKVQGEPPPTG

GA -----
GB1 -----
GA-AB -----
GB1-IGG -----
GA-YSIRK -----
GB1-YSIRK -----
AMBP AVVRTFKDLVTGEVIGFEPTRYTGNIPLSEDGSKDYTNVLGNKYDVSNDPVDLVKEVNGE

GA -----KELKQYGI-G-----DY-----
GB1 -----M-----TY-----
GA-AB -----KELKEAGIKS-----QF-----
GB1-IGG -----RELDKYGV-S-----DY-----
GA-YSIRK -----REFDKYGV-S-----DY-----
GB1-YSIRK -----AEVKAAGFTD-----QK-----
AMBP EYILADIPAENTKGTLSVTKTRARDLYSEEEELKAKGINGSFAFVTPAEYDYVKKTKVEEVN

GA -----YIKL-----
GB1 -----KL-----
GA-AB -----FFNL-----
GB1-IGG -----YKNL-----
GA-YSIRK -----YKKL-----
GB1-YSIRK -----YVDW-----
AMBP RTIKFVYADNVAGLAGTEVFPSQKQTVSYTGSIKLTAEGKAVINSNDRPVYINWKGTGQ

GA -----INNAKTV-----
GB1 -----ILNGKTL-----
GA-AB -----INNAKTV-----
GB1-IGG -----INNAKTV-----
GA-YSIRK -----VNSAKTV-----
GB1-YSIRK -----INAQETV-----
AMBP STDLPPELAVPQKEGYIASVEKVVPVQATTATDEDYEYVVKYTAIQKAKTIFVDEKGNAIPG

```

Figure 43. Continued.

```

GA -----
GB1 -----
GA-AB -----
GB1-IGG -----
GA-YSIRK -----
GB1-YSIRK -----
AMBP VAEITEQGGSETPLTKEADV KAKI KELENKGYELVSNTYPEGGKFDTDKD TDQEFKVILK

GA -----
GB1 -----
GA-AB -----
GB1-IGG -----
GA-YSIRK -----
GB1-YSIRK -----
AMBP QKEVTVPDQPKTPGTPVDLNNPDGPKYPAGLEEKDLNKT VTRTITYVYEDGTPVLNEDG

GA -----EGVESLKNEILKA-----
GB1 -----
GA-AB -----EGVESLKNEILKAHASRSATVDDIKAGD
GB1-IGG -----EGVKALIDEILAA-----
GA-YSIRK -----AGVKKLQAQVVES-----
GB1-YSIRK -----QDVNGVKREILTT-----
AMBP TPKTVTQEAKFTREAKVNLVTGEV TYGDWSEAKDLAEVKSPVVTGFLADKASVPVNVVTG

GA -----
GB1 -----
GA-AB TKVTGTGVPGATIYVTKLPK-----
GB1-IGG -----LPK-----
GA-YSIRK -----
GB1-YSIRK -----VPD-----
AMBP DSKDITEVV TYKPIGSWIPNIPGQPTNPIKY PNNPDDPTQPGKPTEVLPYVPGFTPKDKD

GA -----
GB1 -----
GA-AB -----NGVRDGSSSSSATVGEDGNWSVNLTEP
GB1-IGG -----
GA-YSIRK -----
GB1-YSIRK -----E-----EQSGIEGETSVDS
AMBP GNPLKPVDPADPTKGYIVPDLPTDPSQDTPIN YVKDTQKAKTTFFVDEKGNPIPGVDAITE

GA -----
GB1 -----KGETTTEAVDAATAEKVF-----
GA-AB AEEGDRFSIIQVEPGKGDSKSVIKVVENAE EKPEIQDGFDT EEEAIAAAKKALENDQIN-
GB1-IGG -----TDYKLI LNKGTLKGETTTEAVDAATAEKVF-----
GA-YSIRK -----
GB1-YSIRK EYDFNEPEIDWSGNDEAVAE EEAQEEVYTLNYYAQRTQGQNGATT VKASSPREALEYF-
AMBP EGDSDTPLTKEADV KAKI KELENKGYELVSNTYPEGGKF DDKD TDQEFKVT LKAKEVTV

```

Figure 43. Continued.

```

GA -----
GB1 -----
GA-AB -----
GB1-IGG -----
GA-YSIRK -----
GB1-YSIRK -----
AMBP TPDQPKTPGTPVDPNPNPDGPKYPAGLEEKDLNKTVTRTITYVYADGTPVLNEDGTPKTVT

GA -----
GB1 --KQYANDNGVDG-----
GA-AB --KSYTINQGADGKYYYVLSPVENDEEEKPEEEKPAEQDGYATYEEAEAAKKALENDPIN
GB1-IGG --KQYANDNGVDG---EWTYDDATKTFTVTEKPEVIDASELT-----
GA-YSIRK -----
GB1-YSIRK --QAFLNENGLDAADFNSYDSESRAFTASEKIEGEASVDSEYDFIRPDFDWSGLEEAED
AMBP QEAKFTREAKVNLVTGEVTYGDWSEAKDLPEVKSPVVKGYLADKATVPATKVTADSENTK

GA -----
GB1 -----
GA-AB KSY-----TISQGANGRYYYYLLSPNPAETPEKPEE-
GB1-IGG -----PAVTTYKLVINGKTLKGETTTKA
GA-YSIRK -----
GB1-YSIRK D-----EEVQEEIYTFVYIIQNTKGKNGATTVKA
AMBP EVVTYKPIGSWIPNIPGQPTNPIKYPNDPTDPTKPGQPTETLPYVPGFTPEDKDGNNPLKP

GA -----
GB1 -----
GA-AB -EKPEAQDGYA-----
GB1-IGG VDAETAEKAFK-----
GA-YSIRK -----
GB1-YSIRK SSPEEAKAYFE-----
AMBP VDPNDPTKGYEVPSIPTNPGEDTLINYVANKANLVVKYVDENGKELLPTETKEGKVGDDY

GA -----
GB1 -----EWTY-----
GA-AB -----TYEEAEAAAKEALKN-----
GB1-IGG -----QYANDNGVDG-VWTY-----
GA-YSIRK -----
GB1-YSIRK -----EFAKENDLGELDWTY-----
AMBP STSGKVITGYVLDRVEGEAKGKIGTDGTTVTTYVYKPLGSWIPNIPGQPTNPIKYPNDPTD

GA -----
GB1 -----
GA-AB -----
GB1-IGG -----
GA-YSIRK -----
GB1-YSIRK -----
AMBP PTKPGQPTETLPYVPGFTPEDKDGNNPLKPVDPNDPTKGYEVPSIPTNPGEDTPINYVANK

```

Figure 43. Continued.

```

GA -----LPTE-----
GB1 -----DDATKTFTVTE-----
GA-AB -----DKINKSYSIRQGADGRYYYYVLSPEAETPSTPEVTPTP-----GVTPT
GB1-IGG -----DDATKTFTVTE-----MVTEVPGDAPTEPEKPEASI-----
GA-YSIRK -----AK-----
GB1-YSIRK -----DEDTKTFTARE-----KVEESQTIEGETSVDSVYDFNRPEIDWSGAEEV
AMBP ANLVVKYVDENGKELLPTETKEGKVGDDYSTSGKVITGYVLDRVEGEAKGKIGTDGTTVT

GA -----
GB1 -----
GA-AB PGGQPSTPEVPYTPYVPSTPGKEDKKPGEDKKPEDKKP-----GEDKKPEDKKPGE
GB1-IGG -----PLVPLTPATPIAKDDAKKDDTKKEDAKK-----PEAKKEDAKKAET
GA-YSIRK -----
GB1-YSIRK YERELQATKEAAIAELQGIYIKDEELFGRIQDAERIEE-----VKKLRSDAIDARQ
AMBP YVYKPLGSWIPNIPGQPTTPIKYPNDPQDPTKPGQPTEVLPYVPGFTPEDKDG NPLKPVD

GA -----
GB1 -----
GA-AB DKKPGKE-----
GB1-IGG LPTTGEG-----
GA-YSIRK -----
GB1-YSIRK ALVGKLD-----
AMBP PKDPSKGYVVPNIPTNPGEDTPINYIPNVTPNGDQDGYTPQPKPQPEQVVYYVDENGKD

GA -----
GB1 -----
GA-AB -----
GB1-IGG -----
GA-YSIRK -----
GB1-YSIRK -----
AMBP IAPSEKGAQAPKGISGYEYVTTTKDPNGNLVHHYKKVATPQPVPTTPETPEQPVAPVQPE

GA -----
GB1 -----
GA-AB -----EKPANPAKPAKEEKEKTDSPNKKKKLKPAGSEAEILTLAAAALSATAG
GB1-IGG -----SNP-----FFTAAALAVMAGAG
GA-YSIRK -----
GB1-YSIRK -----DQEA-----SADSEYDFVRPDVD
AMBP QPTTPTQPAVPTPAETSVPTDSATKPATPKYVDGQKELPNTGTEANASLAAFGLLGALGG

GA -----
GB1 -----
GA-AB AFVSLKKRK--
GB1-IGG ALAVASKRKED
GA-YSIRK -----
GB1-YSIRK LTGLDEGYARE
AMBP FGLLARKKKED

```

Figure 43. Continued.

		10	20	30	40	50
GA	MEAVDANSLA	AFEPALFELK	IGIGDYY	FLINNAKTVE	GVESLKN	EIFALPTE
GB1	MTYKLI	NGKTLKGETT	TEAVDAATAEKV	FKQYANDNGVD	GEWTYDDATKT	FTVTE

Figure 44. Condensed GB1 and GA sequences from Figure 42. Residues common to all related sequences from Figure 43 are indicated in red. Residues common to only the GA branch are indicated in green and those common to the GB1 branch alone are indicated in purple. Highlighted in yellow are positions that are similar in identity or character within their evolutionary lineage.

The YSIRK proteins in our study are characterized by the LPxTG or YSIRK signal motif only. There is very little information about their structure and function. Since the YSIRK signal was central to determining a possible ancestor we assessed the presence and evolutionary conservation of motifs relevant to our study (Table 5). The protein families database (pfam) (27.0) contains a large collection of protein families already assessed and represented by sequence alignments and hidden Markov models and is powerful in searching for motifs and domains present in a query sequence [392, 393]. Analyzing the mucus-binding protein using pfam revealed the YSIRK signaling and LPxTG motif. The LPxTG motif is a cell wall targeting sequence that is mediated by transpeptidase sortase resulting in peptide bond anchoring to cell wall peptidoglycans [394]. This motif is significantly useful for binding cell surface proteins to the outer cell wall of gram-positive bacteria. Interestingly, the LPxTG motif is conserved through evolution to the full length proteins of GB1 and GA. However, the YSIRK signaling motif is only conserved in the GB1 related YSIRK protein. It is likely that the GA related

YSIRK protein, which is only a partial sequence, is missing the motif due to missing residues. Although it is missing the YSIRK signal motif, it does however contain the GA module domain which is not seen in the mucus-binding protein. This suggests that the GA module may have developed early in the evolutionary process. In comparison, the IgG-binding domain was only present in the IgG-binding protein indicating that this domain may have evolved after the GA module in evolution. Taken together, these results seem to indicate that the GA module may be older than IgG-binding domain in terms of evolution

Table 5. Motifs found in GB1 and GA related proteins using pfam analysis

Protein	Motif(s)	Alignment Region		E-value
		Start	End	
IgG-Binding Protein G	GA Module	68	118	8.7E-14
		145	193	4.9E-12
	IgG-Binding	202	256	9.8E-33
		272	326	5.6E-33
	LPXTG Motif [LPxTG-Hydrophobic domain- Positive Residue(s)]	381	418	1.2E-10
Albumin-Binding Protein	GA Module	64	116	8.40E-07
		169	219	1.30E-12
		279	331	2.90E-12
		383	439	2.00E-11
		514	558	3.20E-09
	LPXTG Motif [LPxTG-Hydrophobic domain- Positive Residue(s)]	915	954	1.60E-06
(GA) YSIRK, Partial	GA Module	105	154	7.8E-13
		182	229	2.3E-10
(GB1) YSIRK	YSIRK type signal peptide [YF]SIRKxxxGxxS[VIA]	16	41	8.30E-12
Mucus-binding Protein	YSIRK type signal peptide [YF]SIRKxxxGxxS[VIA]	14	36	7.40E-09
	LPXTG Motif LPxTG-Hydrophobic domain-Positive Residue(s)	1893	1928	1.10E-07

As far as we could see the GA and GB1 related YSIRK proteins are functionally and structurally uncharacterized. In addition, there is very little structural information on mucus-binding proteins. To investigate the structures of the YSIRK proteins related to GB1 and GA we used the SWISS-MODEL and I-TASSER modeling programs. SWISS-MODEL is a homology modeling program that builds a model based on a database of

already solved structures [395]. The SWISS-MODEL of the GA-Related YSIRK protein revealed a 3-helical bundle topology very similar to the WT-GA structure (Figure 45A). On the other hand the GB1-Related YSIRK protein model shows what appears to be two partially formed $4\beta+\alpha$ domains similar to that found in WT-GB1 (Figure 45B). These models further provide supporting evidence that they could indeed be distantly related to GA and GB1, respectively. The SWISS-MODEL of the ancestral mucus-binding protein has an interesting structure and appears to have a primitive form of the WT-GB1 protein on the C-terminal end. This could indicate that the mucus-binding proteins are more closely related to the IgG-binding domain structure. Interestingly, this does not support our phylogenetic analysis that revealed a greater relatedness to WT-GA.

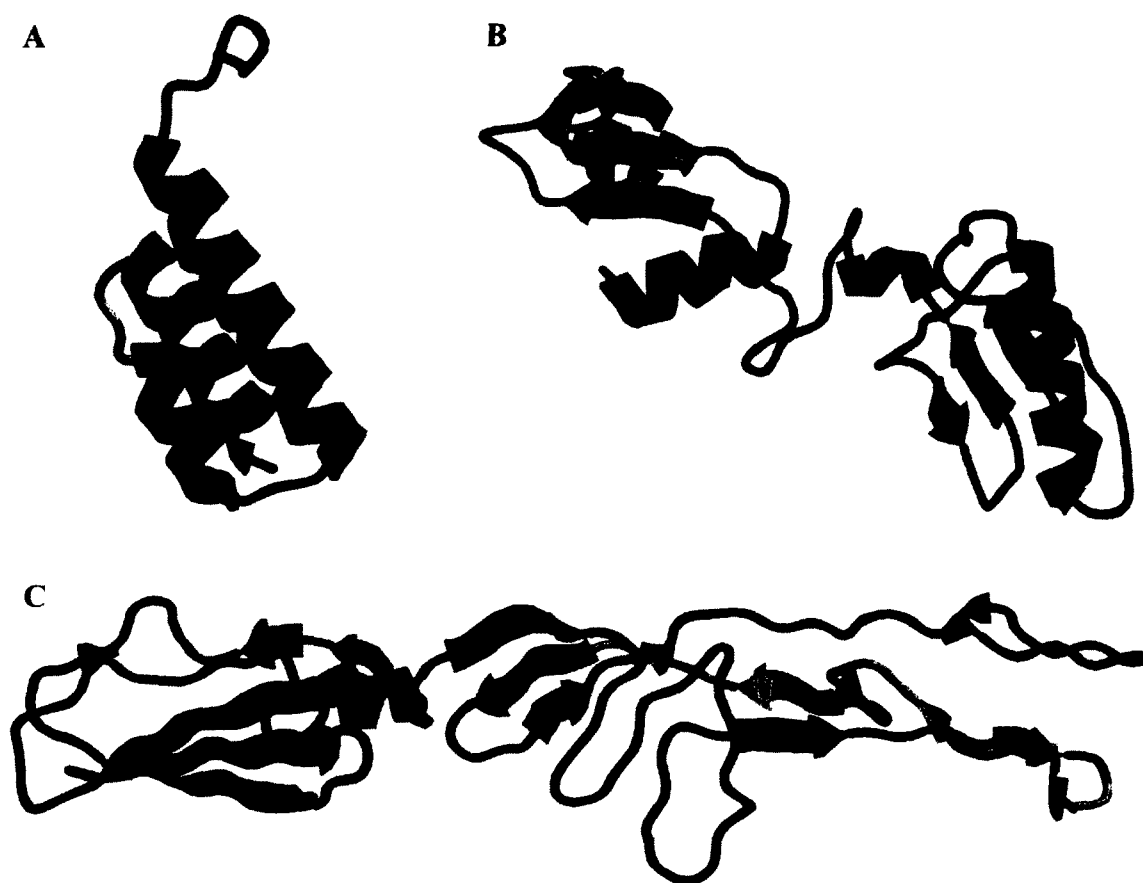


Figure 45. SWISS-MODEL of GA and GB1 evolutionarily related proteins.

Structures were generated using SWISS-MODEL [395]. Model of the (A) GA-Related YSIRK protein, (B) GB1-Related YSIRK protein and (C) evolutionary mucus-binding protein ancestor. Structures are colored from the N-terminal in red to the C-terminal in blue through the sequence of colors of the visual electromagnetic spectrum. Images were created using RasMol (Ver. 2.7.2.1.1).

We further assessed the structures using the I-TASSER modeling program which uses sequence homology threading algorithm that does not rely on any evolutionary information [102, 396]. The limitations in I-TASSER maximum protein length prevented us from modeling of the ancestral mucus-binding protein. However, structures of the GA and GB1 related YSIRK proteins were constructed (Figure 46). There were some significant differences in the structures obtained from I-TASSER in comparison to the SWISS-MODEL. The GA related YSIRK protein seems to still have a 3-helical bundle structure although now it is extended through a larger portion of the protein sequence (Figure 46A). The lack of secondary structure continuity throughout the whole structure is something that would be expected in an ancestral protein. In the GB1 related YSIRK protein the chain seems to have grouped into four folded domains of predominately β -sheet content (Figure 46B). The N-terminal domain appears to have an early resemblance to the WT-GB1 structure with a partially formed α -helix and three β -strands. These structures seem to have predicted structures that resemble their evolutionary progeny in both modeling simulations. Interestingly, an investigation to characterize the crystal structure of a mucus-binding protein (PDB code: 3I57) and characterization of its function revealed a link to Ig-binding function [397]. This is strong evidence, at least in part that the mucus-binding proteins are related to WT-GB1.

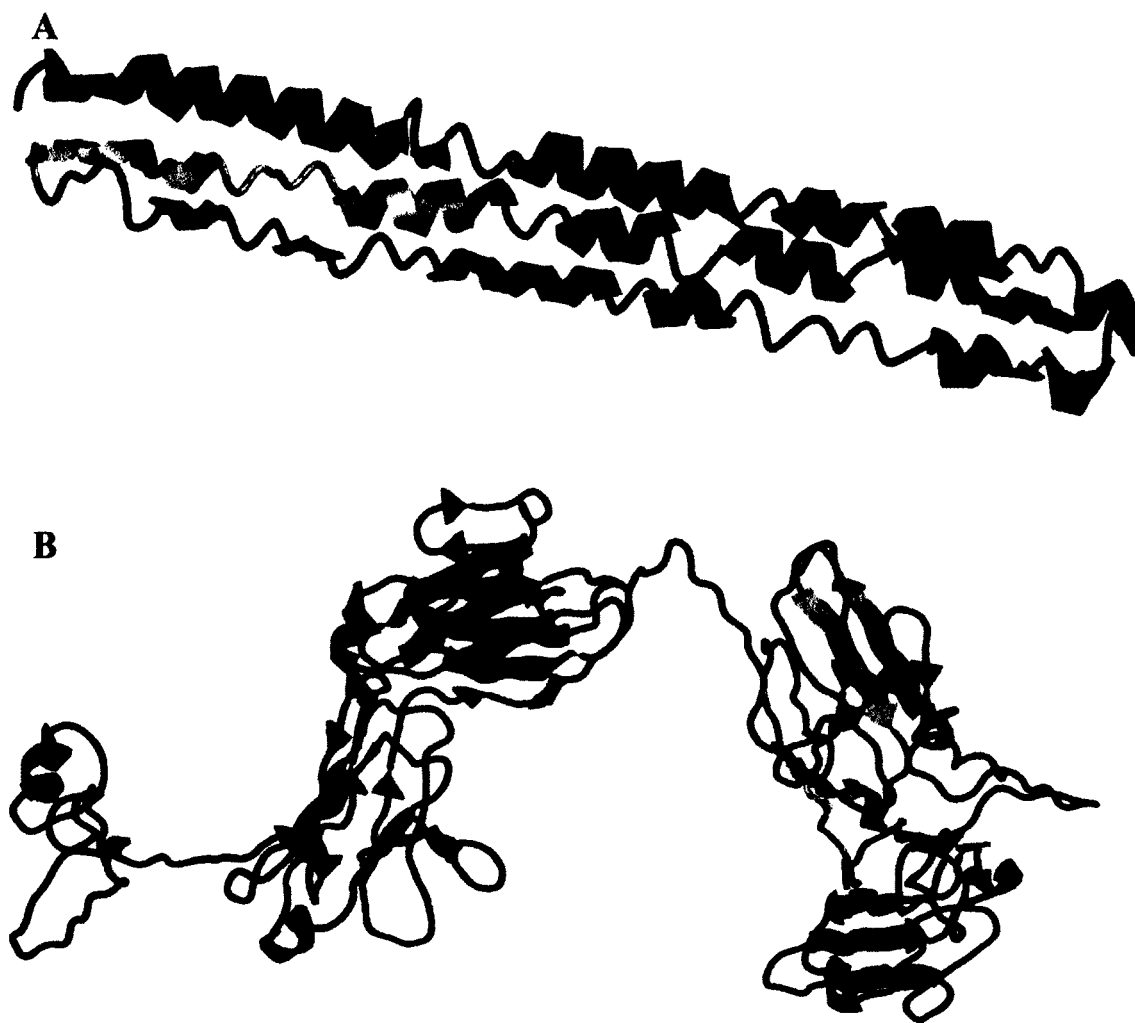


Figure 46. I-TASSER models of GA and GB1 evolutionarily related proteins.

Structures of the (A) GA-Related YSIRK protein and (B) GB1-Related YSIRK protein.

Structures are colored from the N-terminal in red to the C-terminal in blue through the sequence of colors of the visual electromagnetic spectrum. Images were created using RasMol (Ver. 2.7.2.1.1).

It is interesting that we were able to relate both proteins to a mucus-binding type protein. Mucus-binding proteins are believed to help bacteria survive inside a host organism by allowing it to bind to the mucus layer of epithelial cells [41]. This function prevents the bacteria from being expelled from the organism. It could be possible that the albumin-binding and IgG-binding function of GA and GB1, respectively evolved from mucus-binding proteins. The location of both of these types of ligands can be found in the blood of an organism. In the blood the most hazardous enemy to bacteria is the immune system response. It seems that bacteria found in the plasma of an organism may have evolved the mucus-binding protein to counter the immune system by binding immunoglobulins, essentially camouflaging itself in the blood. In addition, the development of the albumin-binding protein would further enhance camouflaging and increase survival of the bacteria in the host organism.

MATERIAL AND METHODS

PSI-BLAST and MUSCLE

To search for and determine a possible ancestral link between the GB1 and GA module, PSI-BLAST was used to search for common sequence identities [362-364]. PSI-BLAST is a method for determining protein sequences with distant similarity to your target sequence. It calculates a position-specific scoring matrix sequence which captures the conservation pattern in an alignment and stores it as a matrix of scores for each position in the alignment. In the position-specific scoring matrix sequence, highly conserved positions receive high scores and weakly conserved positions receive scores near zero. Using this profile in place of the original substitution matrix a BLAST database search is done to detect sequences that match the conservation pattern specified. By finding distantly related sequences in both GB1 (PDB code: 1PGB) and GA (PDB code: 2FS1), sequences found in both PSI-BLAST results suggest a possible divergence from an ancestral protein.

MUSCLE is a multiple sequence alignment program that use algorithms like fast distance estimation using *k*mer counting, progressive alignment using a profile function called log-expectation score and refinement using tree-dependent restriction partitioning (<http://www.ebi.ac.uk/Tools/msa/muscle/>) [398]. The distance estimation is performed using a *k*mer or contiguous subsequence that has length *k*. Sequences that are related tend to have high more *k*mer sequences in common that could be randomly achieved. Positions of identity or similarity is considered significant when the statistical e-value is <0.005.

Determination of Random Sequence Identity

Random small globular proteins were selected by a PDB search for “small proteins” (Figure A3). Resulting proteins were sorted by size and selected based on similar sized but uniqueness of function and species. The 7 proteins were selected and aligned in concert and individually with GA and GB1 using MUSCLE.

Phylogenetic Analysis and Motif Analysis

Phylogenetic analysis and tree construction was done using MEGA version 6 [399]. FASTA sequence files were constructed containing all proteins and imported into MEGA. An internal MUSCLE alignment was computed. The resulting alignment was used to construct the phylogenetic tree. Phylogenetic trees were constructed using statistical methods of Maximum Likelihood or Neighbor-joining with a *p*-distance or Poisson model. Phylogenetic results were tested and verified using either Bootstrap or Interior-branch method. Significant motifs were assessed using pfam database (<http://pfam.xfam.org/>). Each protein was assessed and resulting motifs were provided.

Protein Modeling

The proposed evolutionary related sequence of interest that did not have a high-resolution NMR or crystal structure already assigned from experimental work was modelled using computational methods. The homology modelling programs used were the SWISS-MODEL and I-TASSER [101, 395, 400]. The SWISS-MODEL is a fully automated protein structure homology-modelling server, which can be accessed via the ExPASy web server (<http://swissmodel.expasy.org/>). I-TASSER employs a multiple-threading homology method that builds a model by collecting high-scoring target-to-

template alignments from 9 locally-installed threading programs

(<http://zhanglab.ccmb.med.umich.edu/I-TASSER/>).

CHAPTER III

THE PROPOSED ROLE OF CONSERVED RESIDUES IN THE STABILITY, STRUCTURE AND FOLDING OF THE B1 DOMAIN OF PROTEIN G BY SITE-DIRECTED MUTAGENESIS AND PHI-VALUE ANALYSIS

OVERVIEW

Investigations into the protein folding problem have been significantly furthered by technological advances in protein engineering experiments through the advent of PCR and site-directed mutagenesis. Folding from the ensemble of denatured states by restriction of conformational space to form the initial native-like topology, prior to native-state secondary interaction stabilization, is believed to be due to the formation of an evolutionarily conserved set of amino acids for folding. Residues are typically conserved in a family of proteins because they make critical interactions that are more important in maintaining the fold which could lead to residues clustering together in a hydrophobic core to stabilize the initial native-like topology [401, 402]. This network of conserved amino acids has been the target of a large amount of computational and experimental research which investigate the link between conserved amino acids and how they facilitate rapid and correct folding of a protein in to its native state [37, 137, 359, 360, 402-414]. With the application of site-directed mutagenesis the ability to distinguish the effects of specific amino acids on protein stability and folding kinetics, has led to investigations into the mechanisms of protein folding based on computational

conservation analysis. This type of conserved amino acid view expands the folding nucleus concept of protein folding and there is a growing body of protein engineering work done that shows that these conserved amino acids are preferentially formed in the transition state [15, 35, 415-424]. However, this area is still controversial as there are a few studies that seem to argue against the conservation of a folding nucleus [425, 426]. To further advance our understanding, investigation into the determinants of protein structure and folding from an evolutionary perspective is still an evolving avenue of research as promising new approaches continue to emerge [427, 428].

The cloning of DNA and protein engineering are the basic methodologies in biochemical and biological chemistry and essential in many applications. PCR is an extremely powerful technology, developed in the 1980s by Kary Mullis, and completely changed the biological field [429-431]. Its influence has enhanced the study of protein structure, stability and folding in speed, quantitation and specificity. It has dramatically changed the way we conduct molecular studies and has expanded the possibilities of molecular biology. PCR is used in proteins synthesis to amplify genes for insertion into vectors for controlled expressions and rapidly producing mutations in a process of site-directed primer design [432]. These tasks were much more arduous, requiring the use of recombinant DNA methods. PCR enhanced the identification of novel genes and pathogens and improved quantification of characterized nucleotide sequences [433]. The combination of PCR methods for rapid mutagenesis of genes in clonal DNA (cDNA) is of particular importance to this work. Using site-direct mutagenesis to alter conserved amino acids can lead to a quantifiable decrease in protein stability and cause alterations in secondary and tertiary structure. In addition there can be a decrease in the real-time

folding kinetics as the formation of core native-like interactions can be impaired. Thus, mutagenesis is quite valuable in probing the structure of proteins to discern the importance of specific amino acids to the fold.

Our model system is the streptococcal immunoglobulin-binding domain of protein G with a Thr2Gln mutation to prevent methionine excision. Protein G is a bacterial membrane protein and its function is believed to be related to the survival of the organism. GB1 has been shown to bind the immunoglobulin constant region which appears to provide the bacteria with a immune system camouflage. GB1 is an ideal model because it is a small 56-residue protein whose general folding behavior, stability and 3D structure have been characterized. It can also be refolded at high concentration which is relevant to the studies in Chapter IV. Its structure is made up of two anti-parallel β -hairpins with a stacked α -helix. GB1's comparatively simple structure makes investigating long-range interactions and how they relate to the formation of the tertiary structure visually more simplified. In this chapter we use computational methods to determine the conserved amino acids in GB1. We then experimentally investigate the effects on stability, structure and folding of a conserved residue at position 52. A Phe52Tyr variant was synthesized using site-directed mutagenesis. We then compared the WT and Phe52Tyr variant stability using equilibrium fluorescence, the structure using CD and folding dynamics with stopped-flow fluorescence experiments.

RESULTS AND DISCUSSION

To determine the evolutionary conservation of amino acids in the sequence of GB1 we developed a structural superfamily using the Dali server [434-436]. We uploaded the pdb file of GB1 into the server and obtained a list of proteins whose structures are superimposable with GB1 within a certain degree of error dictated by the intrinsic algorithms. From the server, we selected 13 proteins whose fold matched GB1 (Table 6). To ensure that the structural alignment would provide information on which amino acids were important in the fold and not biological function, we selected a group of proteins whose function varied from that of GB1. In addition we assessed the sequence identity to ensure that it was below 25% for all the select proteins which is considered the “twilight zone” (Table 7) [437]. The “twilight zone” is a percent of identity in which you cannot be sure or guarantee the proteins have the same 3D structure. Thus we work in this region to enhance sequence variability but use known 3D structures. Although the structural alignment provided by the Dali server is quite advanced, it is still not perfect and only provides a preliminary sequence alignment. To verify the sequence alignment provided by Dali, a manual observation of the superimposable structures is required. The side-chain orientation of each amino acid aligned with GB1 in each protein was verified in comparison to that of GB1 using the 3D structure visualization program, RasMol (Figure 47) [438-441]. This is a significant undertaking and was done in collaboration with John Bedford (Graduate Student, Old Dominion University). Once each side-chain orientation was verified the finalized structure-based sequence alignment was completed (Figure 48). As expected most variability in side chain orientation is found in the loop regions due to high variability in the length and stability due to the absence of significant

stabilizing interactions. Also of note is that the third β -strand of GB1 appears to only have one position in which there was total side-chain orientation alignment. This could indicate that the formation of that strand is not evolutionarily conserved for this fold and could also mean that stabilization of this strand is formed post initial collapse of the structure during folding.

Table 6. Proteins selected for structural alignment with GB1

PDB Code	Species	Amino Acid Length	Classification/Function
1PGB	Bacteria (<i>Streptococcus sp. Group G</i>)	56	Immunoglobulin Binding Protein
2PTL	Bacteria (<i>Peptostreptococcus magnus</i>)	78	Protein Binding (Immunoglobulin L Chain)
1RLF	Mouse (<i>Mus musculus</i>)	90	Signal Transduction Protein
3POO	Halophile (<i>Haloflex volcanii</i>)	89	Protein-Binding
1ENF	Bacteria (<i>Staphylococcus aureus</i>)	212	Toxin
1FMA	Bacteria (<i>Escherichia coli</i>)	81	Transferase
2K8H	Human African Trypanosomiasis (<i>Trypanosoma brucei</i>)	110	Signaling protein
1F2R	Mouse (<i>Mus musculus</i>)	87	DNA Binding Protein
1EUV	Baker's Yeast (<i>Saccharomyces cerevisiae</i>)	221	Hydrolase
1WM2	Human (<i>Homo sapiens</i>)	78	Protein Transport
3A4R	Mouse (<i>Mus musculus</i>)	79	Transcription
3PT2	Crimean-Congo hemorrhagic fever virus (<i>Bunyaviridae Nairovirus</i>)	187	Hydrolase/protein Binding
1C4P	β -hemolytic Bacteria (<i>Streptococcus equisimilis</i>)	137	Blood Clotting
2BS2	Proteobacteria (<i>Wolinella succinogenes</i>)	660	Oxidoreductase
1WSP	Rat (<i>Rattus norvegicus</i>)	84	Signaling Protein

Table 7. Sequence identity analysis of GB1 and 13 structurally aligned proteins.

1PGB	100%													
2PTL	10.71%	100%												
1RLF	7.14%	2.56%	100%											
3PO0	1.78%	6.41%	10.11%	100%										
1ENF	5.35%	6.41%	4.44%	6.74%	100%									
1FMA	12.50%	6.41%	8.64%	7.40%	4.93%	100%								
2K3H	10.71%	7.69%	6.66%	7.86%	7.27%	8.64%	100%							
1F2R	5.35%	6.41%	4.59%	5.74%	4.59%	8.64%	10.34%	100%						
1EUV	1.78%	8.97%	3.48%	2.32%	6.97%	8.64%	8.13%	5.81%	100%					
1WM2	7.14%	5.12%	7.69%	8.97%	3.84%	1.28%	5.12%	12.82%	5.12%	100%				
3A4R	3.57%	2.56%	8.86%	3.79%	5.06%	5.06%	6.32%	8.86%	10.12%	3.84%	100%			
1C4P	0%	3.84%	4.44%	7.86%	5.83%	11.11%	6.36%	4.59%	5.81%	2.56%	15.18%	100%		
2BS2	5.35%	3.84%	11.11%	4.49%	6.13%	7.40%	6.36%	6.89%	3.48%	3.84%	5.06%	3.64%	100%	
1WSP	3.57%	3.84%	5.95%	5.95%	3.57%	1.23%	5.95%	5.95%	9.52%	5.12%	7.59%	5.95%	3.57%	100%
	1PGB	2PTL	1RLF	3PO0	1ENF	1FMA	2K3H	1F2R	1EUV	1WM2	3A4R	1C4P	2BS2	1WSP

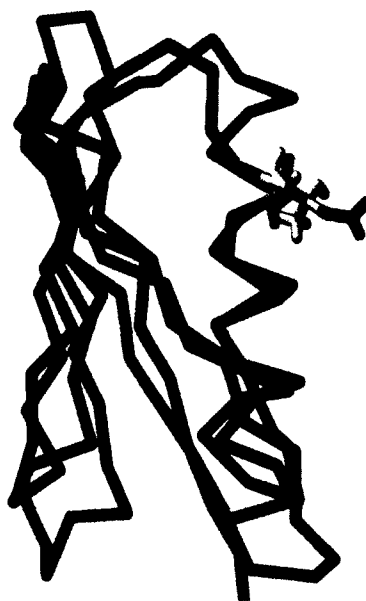


Figure 47. Structural alignment of 1PGB and 2PTL. Image shows an example of a structural alignment and a pair of residues that are aligned in orientation. Residue 32 (green) of GB1 (red) and residue 53 (yellow) of 2PTL (blue) are shown both pointing outward towards solvent.

```

      1      10      20      30      40      50
      |      |      |      |      |      |
[1pgb]-MTYKLIILNGKTLkGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE
[2ptl]-VTIKANLIfagstQTAEFKgTFe-KATSEAYAYADTLkgeYTVDVAdkgYTLNIKF
[1rlf]-RIIRVQMElgdgsVYKSILVT--dkAPSVISRVLkknseFELVQldashDFLLRQ
[3po0]-GSMEWKLF-ADlarTVRVDVDtvGDALDALvgahlesrv-iNVLrN--gdELALFP
[1enf]-RVIGANVWVdiqkETELIRTVtlqELDIKIRKILSDky--GLIEFDMkishIDVNL
[1fma]--MIKVLFFrelvtdATEVAad--fptVEALRQHMAAalalLLAAVn--gdEVAFFP
[2k8h]-VAVKVVNA---dgaEMFFRIKs-rtALKKLIDTYCkkqnsVRFLFd--ddVIDAMV
[1f2r]-KCVKLRLAlh--sackFGVAArsCQELLRKGCVRFq-----sRLCLfpglaELLLLT
[1euv]-INLKVSDg----ssEIFFKIKhttpL-RRLMEAFKRqgkeLRFLYd--ndIEAHR
[1wm2]-INLKVAGQ---dgsVVQFKItplSKLMKAYCERqgl--rqIRFRFd--edTIDVFQ
[3a4r]-LRLRVQgk--ekhqMLEISLSplkVLMSHYEeamgl--hkLSFFfd--gdLIEVWG
[1c4p]-VEYTVQFTpfrpglKDTKLLitsqELLAQAqsilnkpgytYeRSsivtliSEKYYV
[1qla]-RMLTIRVFkYphfqEYKIEeap-smtIFIVLNmirepdlnmMin-lfedGVITLLP
[1wsp]-IVVAYYFcg--epiPYRTLVravGQFKE-LL---tkkg-sYRYFYfkkiIGKVEK

```

Figure 48. Finalized structure-based sequence alignment. All positions were verified visually and any position that was not conserved in side-chain orientation is shown as lowercase letters. Positions in red show positions of complete conservation of side-chain orientation with no more than one gap.

In order to know which amino acids to select for mutagenesis a conservation analysis is performed to determine position specific conservation over the superfamily (Figure 49). Using a modified Shannon's entropy equation amino acid conservation is determined based on the number of amino acid types at each position. From the conservation analysis we found that there were 12 residue positions that were considered evolutionarily conserved: Tyr3, Lys4, Leu5, Thr18, Ala20, Ala26, Phe30, Glu42, Asp46, Lys50, Phe52 and Val54. There are 11 positions that were considered moderately conserved (>0.30) and 1 position, residue Ala26, considered highly conserved (>0.45). It is interesting to note that there is at least one conserved amino acid found in each major secondary structure component.

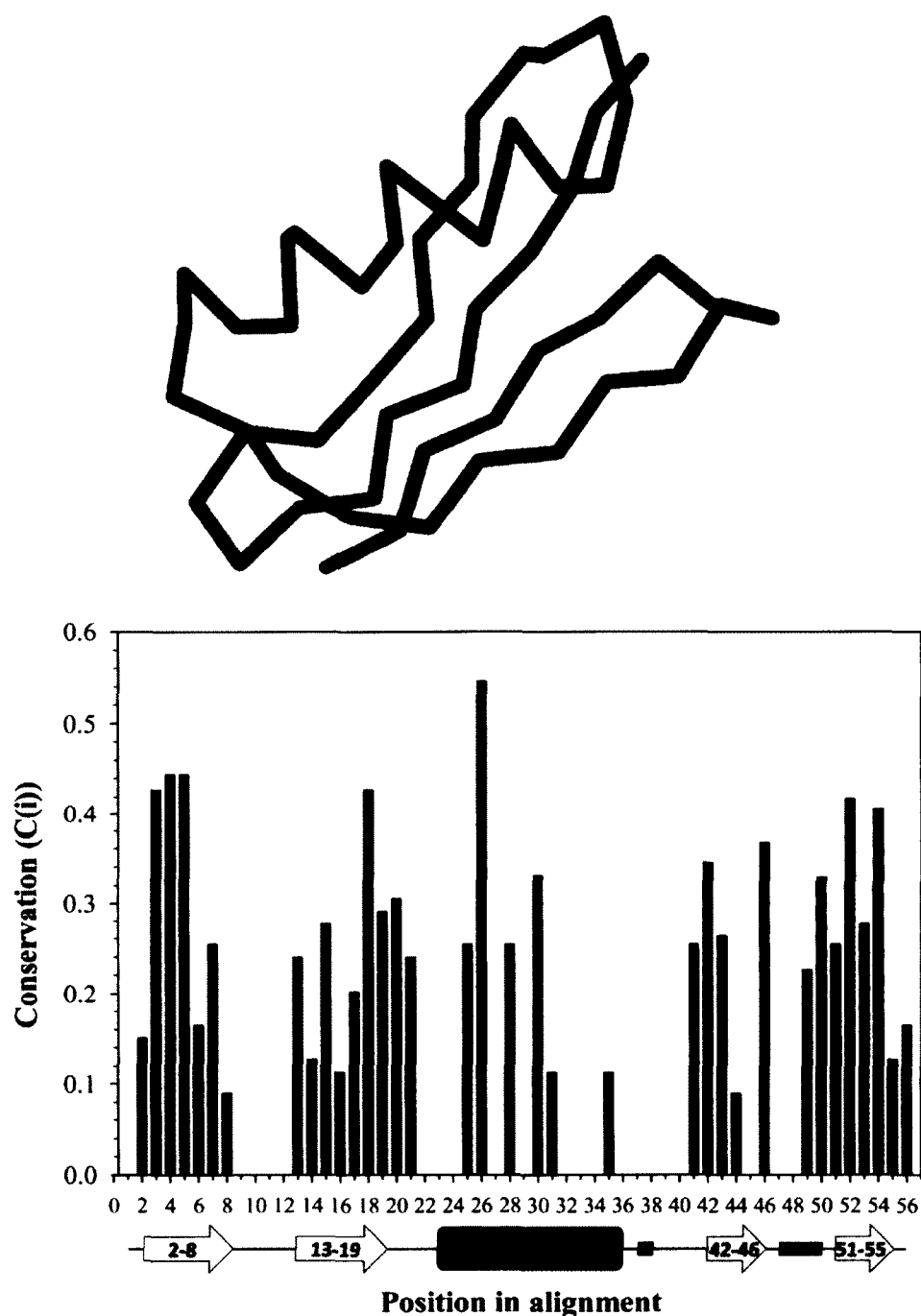


Figure 49. Amino acid position conservation analysis. Positions colored in blue are positions considered conserved. Positions >0.45 are considered highly conserved and $0.45 >$ positions >0.30 are considered moderately conserved. Arrows in yellow indicate β -strands, the rectangle in magenta indicates an α -helix and green rectangles indicated β -turns. Backbone structure was created using RasMol (Ver. 2.7.2.1.1).

The initial analysis determined conservation irrespective of amino acid character thus we also wanted to get a sense of conserved positions with respect to character. The conservation analysis was conducted by counting similar character types rather than the same specific amino acid and the results were plotted similarly (Figure 50). The data indicated that 11 positions were conserved in amino acid character: Tyr3, Leu5, Leu7 Thr18, Ala20, Ala26, Phe30, Gly41, Trp43, Phe52 and Val54. In order to discern less, moderately and highly conserved residues we counted any positions less than 2 standard deviations (StDev) from the mean as less conserved, positions greater than 2 StDev but less than 3 StDev as moderately conserved and any positions greater than 3 StDev as highly conserved. The 11 positions were separated into 4 moderately conserved (Leu7 Thr18, Ala20 and Gly41) and 7 highly conserved positions (Tyr3, Leu5, Ala26, Phe30, Trp43, Phe52 and Val54). Between the two conservation analyses there are 4 positions (Lys4, Glu42, Asp46 and Lys50) that were considered conserved but are not similarly conserved in the character analysis. This indicates that during evolution these positions typically maintained the same amino acid and their role may be more dependent on the particular amino acid structure. However, when these positions are modified it does not favor a particular amino acid character type and this could mean that its role in folding of GB1 could be secondary. Interestingly, the following 8 positions are conserved in both amino acid position and character: Tyr3, Leu5, Thr18, Ala20, Ala26, Phe30, Phe52 and Val54. In addition, 3 positions (Leu 7, Gly41 and Trp43) that were not conserved in amino acid position are now considered conserved with respect to amino acid character. This distinction indicates that these positions may be important in the folding of GB1 because of the character of the amino acid. This could suggest that during evolution when

these positions were varied they did not favor any specific amino acid in particular but required that the character of the position be maintained for folding.

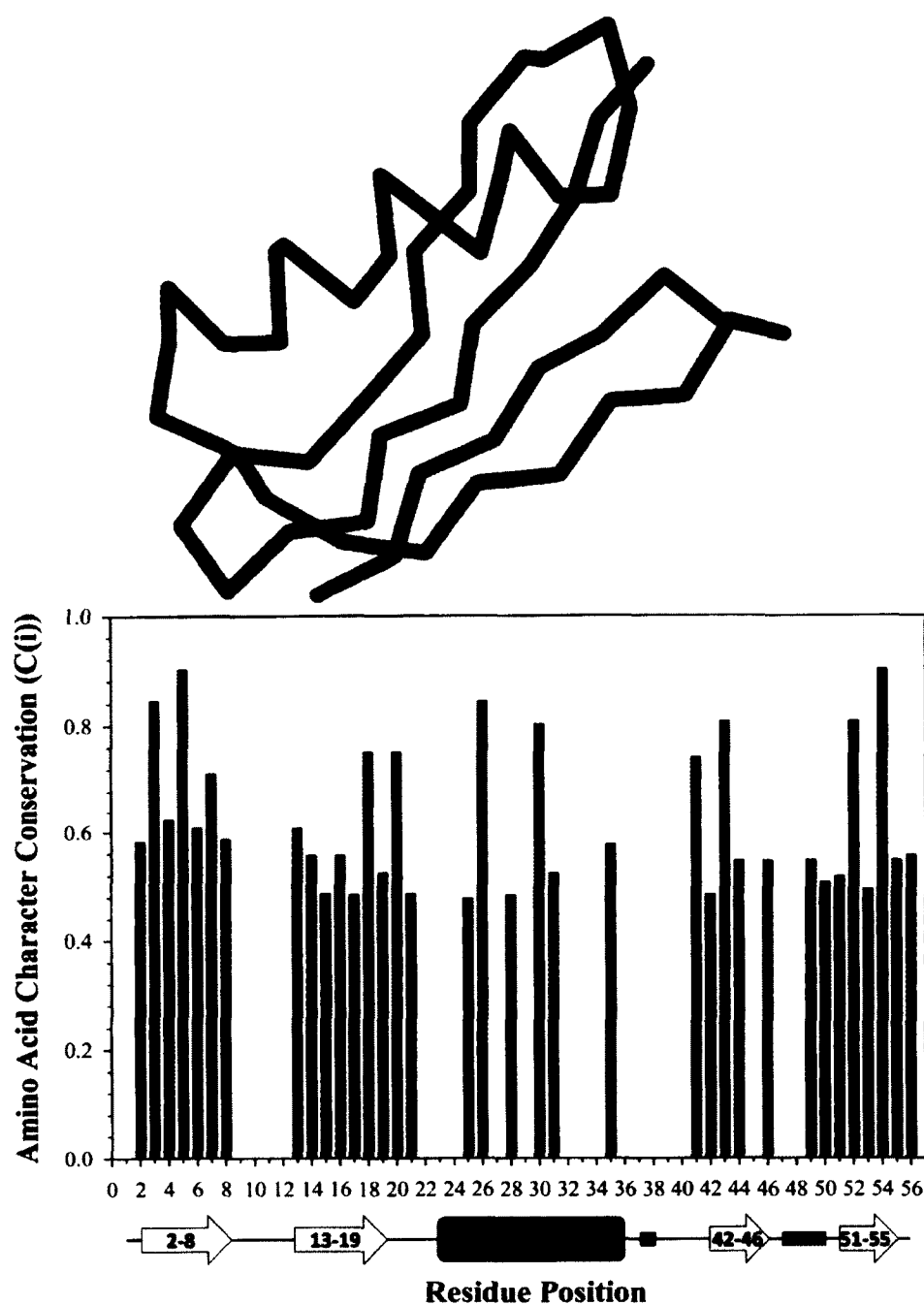


Figure 50. Amino acid character conservation analysis. Positions colored in blue are positions considered moderately conserved whereas those colored in red indicate highly conserved positions. Arrows in yellow indicate β -strands, the rectangle in magenta indicates an α -helix and green rectangles indicated β -turns. Backbone structure was created using RasMol (Ver. 2.7.2.1.1).

To identify hydrophobic positions versus hydrophilic positions a hydropathy analysis was done (Figure 51). From the hydropathy analysis we see that of the 15 positions conserved by amino acid type or character, 11 were hydrophobic while 4 were hydrophilic in nature. This makes sense as the 4 positions considered hydrophilic are either acidic or basic in character and would be expected to be found on the surface of the protein exposed to the solvent. Based on all the bioinformatics information gathered there are 15 positions that were revealed to have conservation in one way or another. A summary of all the data can be found in Table 8.

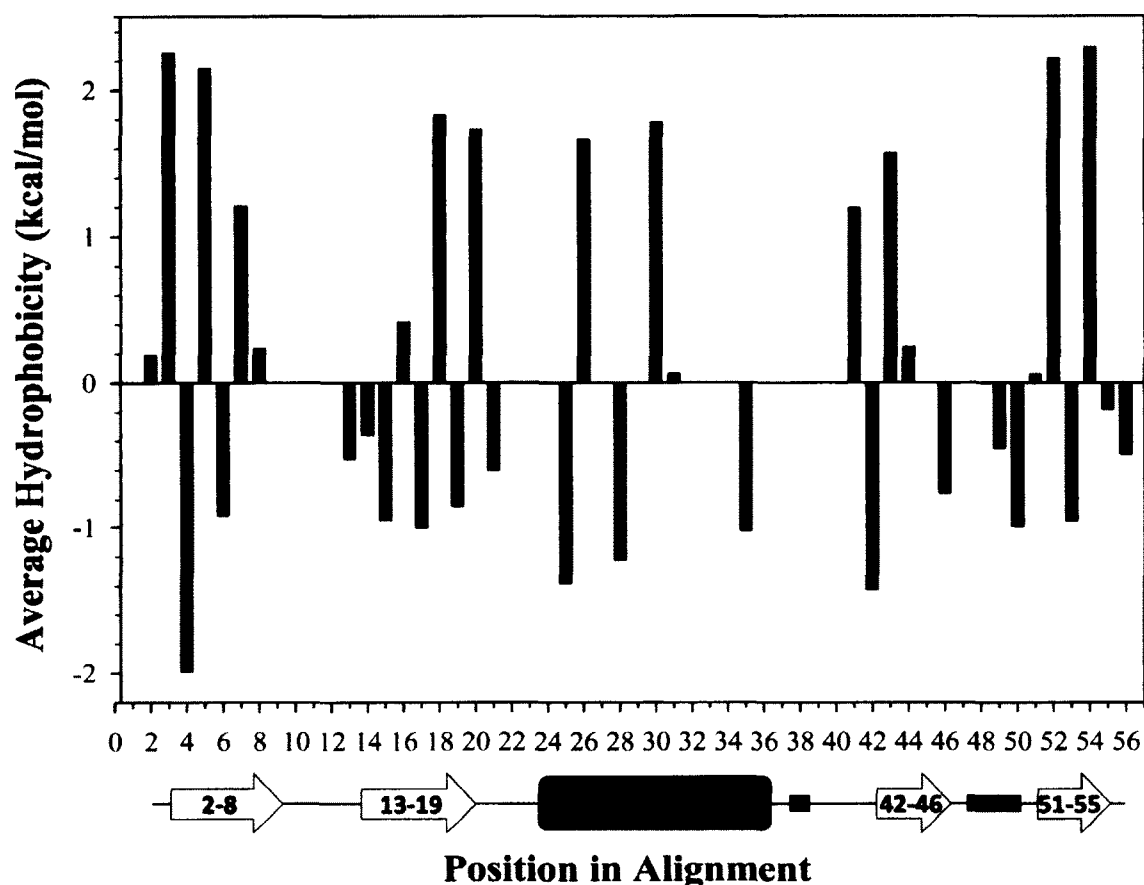


Figure 51. Position specific hydropathy analysis. Positions colored in blue are positions considered conserved either by position or character conservation. Positive values indicate hydrophobicity and negative values indicate hydrophilicity. Arrows in yellow indicate β -strands, the rectangle in magenta indicates an α -helix and green rectangles indicated β -turns.

Table 8. Summary of all conserved positions from the conservation analysis

	Amino Acid (GB1)	Secondary Structure	Side-Chain Orientation	Amino Acid Conservation	Character Conservation	Hydropathy
3	Tyrosine (Y)	β -Strand-1	Aligned	Moderately	Highly	Hydrophobic
4	Lysine (K)	β -Strand-1	Aligned	Moderately	Less	Hydrophilic
5	Leucine (L)	β -Strand-1	Aligned	Moderately	Highly	Hydrophobic
7	Leucine (L)	β -Strand-1	Aligned	Less	Moderately	Hydrophobic
18	Threonine (T)	β -Strand-2	Aligned	Moderately	Moderately	Hydrophobic
20	Alanine (A)	Loop-2	Not Aligned	Moderately	Moderately	Hydrophobic
26	Threonine (T)	α -Helix-1	Not Aligned	Highly	Highly	Hydrophobic
30	Phenylalanine (F)	α -Helix-1	Aligned	Moderately	Highly	Hydrophobic
41	Glycine (G)	Loop-3	Not Aligned	Less	Moderately	Hydrophobic
42	Glutamic Acid (E)	β -Strand-3	Not Aligned	Moderately	Less	Hydrophilic
43	Tryptophan (W)	β -Strand-3	Aligned	Less	Highly	Hydrophobic
46	Aspartic Acid (D)	β -Strand-3	Not Aligned	Moderately	Less	Hydrophilic
50	Lysine (K)	β -Turn-2	Not Aligned	Moderately	Less	Hydrophilic
52	Phenylalanine (F)	β -Strand-4	Aligned	Moderately	Highly	Hydrophobic
54	Valine (V)	β -Strand-4	Aligned	Moderately	Highly	Hydrophobic

To determine the importance and effect the 15 conserved amino acids have on the stability and folding kinetics we can use site-directed mutagenesis to mutate the position to an alanine or glycine. In this chapter, we selected the phenylalanine at position 52 as the first mutation to be studied. We initially selected this residue in order to selectively label a specific long-range interaction to be studied by NMR, which will be discussed in chapter IV. Thus the mutation to tyrosine is much less disruptive. However, important information about the effect on the structure and stability is determined experimentally. The WT and Phe52Tyr variant were both overexpressed in *E. coli* and both proteins were initially observed to be highly stable and non-toxic to bacterium. We were able to overexpress both proteins and produce anywhere from ~200-300 mg of protein per 6 L of

bacterium. This indicates that in both cases GB1 is highly soluble and stable at high concentrations *in vivo*. Any adverse effects of GB1 and its variant is not indicated as evidenced by the high expression levels and due to the absence of inclusion bodies in the post sonication milieu (which typically indicates the protein is not favorable to bacteria).

To determine the structural differences between the WT- and Phe52Tyr-GB1 we used far- and near-UV CD (Figure 52). Far-UV CD of the Phe52Tyr variant indicated that the secondary structure does not change its shape in any significant way. However, there is a decrease in the signal indicating a possible increase in overall secondary structure stability. It has already been discussed that tertiary interactions can stabilize secondary interactions. The tyrosine could be in this case causing an increase in secondary content, perhaps due to an increase in helicity. Similarly, the tertiary structure monitored by near-UV CD shows a similar pattern between WT and Phe52Tyr variant. There seems to be a change in the amino acid environment of the tertiary structure in the core due to the additional -OH group. In both cases, at 100 °C both WT- and Phe52Tyr-GB1 show a random coil secondary structure and a complete loss of tertiary structure indicating a fully unfolding state. Figure 53 shows the thermal unfolding transition for WT-GB1 and the Phe52Tyr variant. It reveals that the variant is less thermostable than the WT protein. The WT protein has a temperature midpoint (T_m) of 80 °C while the Phe52Tyr variant has a T_m of 75 °C. Although we see structural changes associated with this mutation which is suggestive of a potentially slight increase in secondary and tertiary structure content in the CD, this does not correlate well with the stability data.

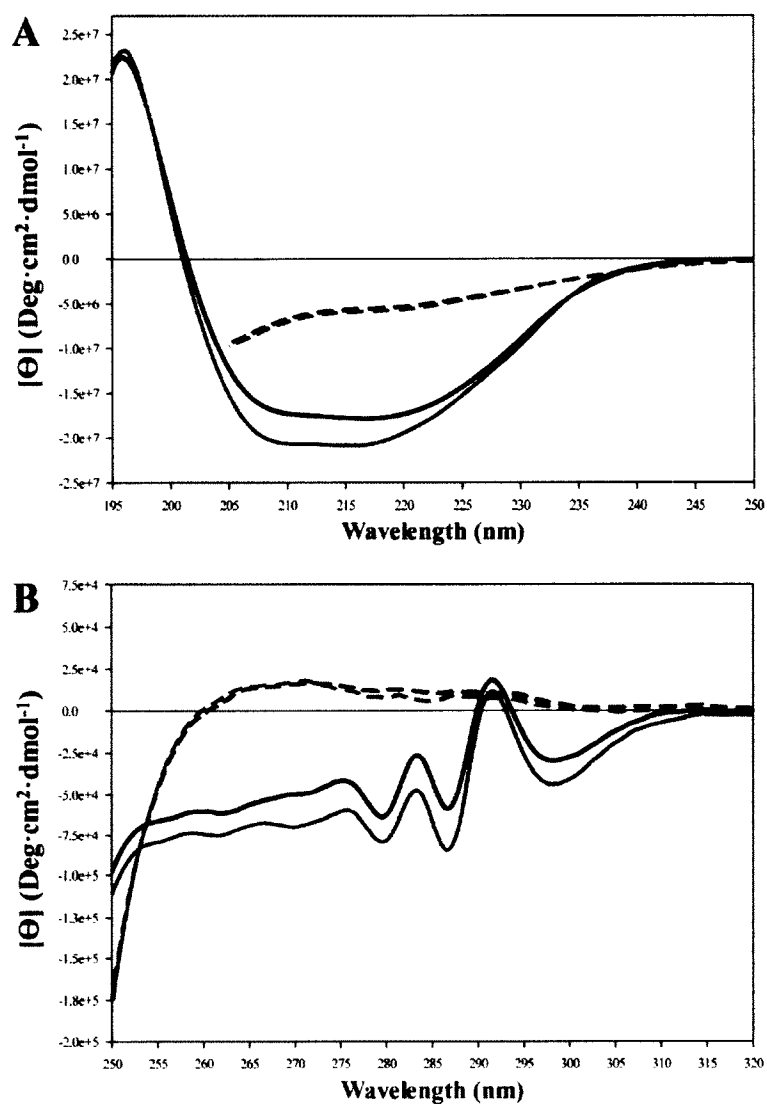


Figure 52. Circular dichroism spectra of WT and Phe52Tyr-GB1. WT-GB1 at 20 °C (Solid-line) and 100 °C (dotted-line). Phe52Tyr-GB1 at 20 °C (dashed-line) and 100 °C (dot-dot-dash-line). **(A)** Far- and **(B)** near-UV CD. The data was plotted and visualized in SigmaPlot (Ver. 12.5)

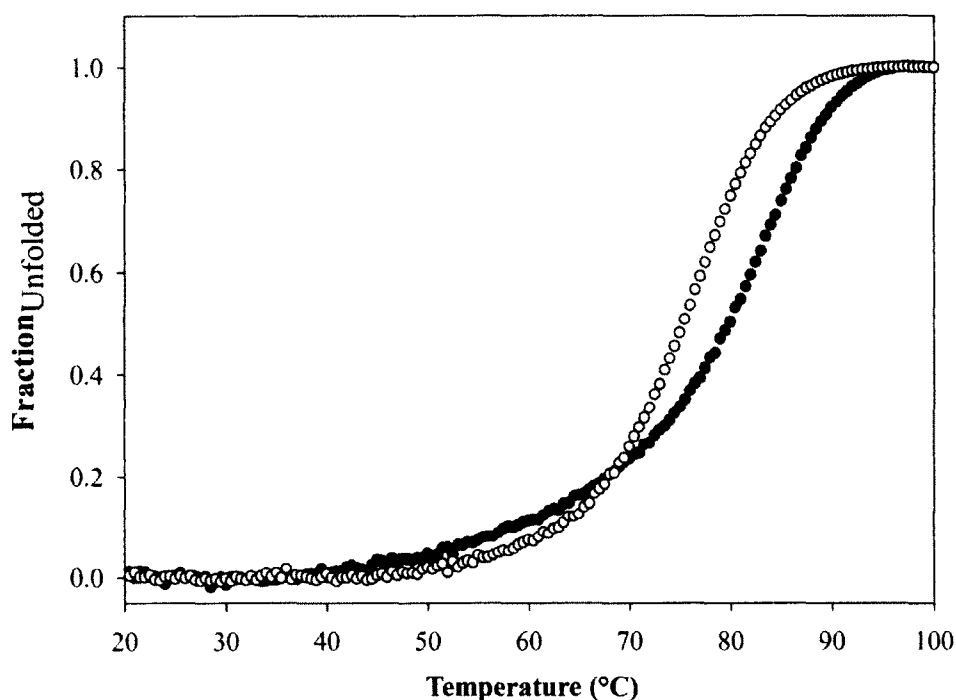


Figure 53. Thermal denaturation of WT- and Phe52Tyr-GB1. WT-GB1 is indicated by the filled circles and Phe52Tyr-GB1 is indicated by the unfilled circles. The data was plotted and visualized in SigmaPlot (Ver. 12.5)

Equilibrium experiments are useful in helping to define the native and unfolded state structures and their relative stability. In addition, the denaturation curves provide a basis for future kinetic experiments by showing how the native state changes as function of denaturant. To determine the stability of WT- and Phe52Tyr-GB1 the folding was studied using intrinsic tryptophan fluorescence in 0.1 M tris base at pH 7.0. Interestingly, the initial baseline of the native state from the fluorescence data is not flat, which is what was expected (Figure 54). There is a decreasing linear change in the fluorescence native baseline, which indicates that the environment of the tryptophan is becoming more solvent exposed as the concentration of guanidine hydrochloride (Gnd-HCl) increases

from 0 to about 1 M. At this point the unfolding transition predominates. This effect we believe is due to the amino acids neighboring the tryptophan. The tryptophan points into the hydrophobic core however, directly on the outer surface of the tryptophan there is a lysine at position 31 that appears to form a cation- π interaction which partially protects the tryptophan from solvent exposure (Figure 55). Cation- π interactions are becoming increasingly recognized as relevant non-covalent binding interactions in the stability of protein structures [442-446]. A computational study indicated that out of 593 proteins lysine is found to interact 30% of the time with tryptophan in a cation- π interaction [442-446]. Upon addition of Gnd-HCl we believe the cation- π becomes increasingly destabilized allowing water to come into contact with the tryptophan. There are examples of proteins that have similar cation- π interactions with little change to the equilibrium baselines such as the human serum retinol-binding protein (PDB code: 1jyd) [447]. This change due to the addition of Gnd-HCl could be unique to GB1 due to how close Trp43 is to the outer surface of GB1 or the angle of interaction. In addition, GB1 only has one tryptophan providing the average fluorescence signal, whereas in other proteins the fluorescence contribution may come from multiple buried tryptophans.

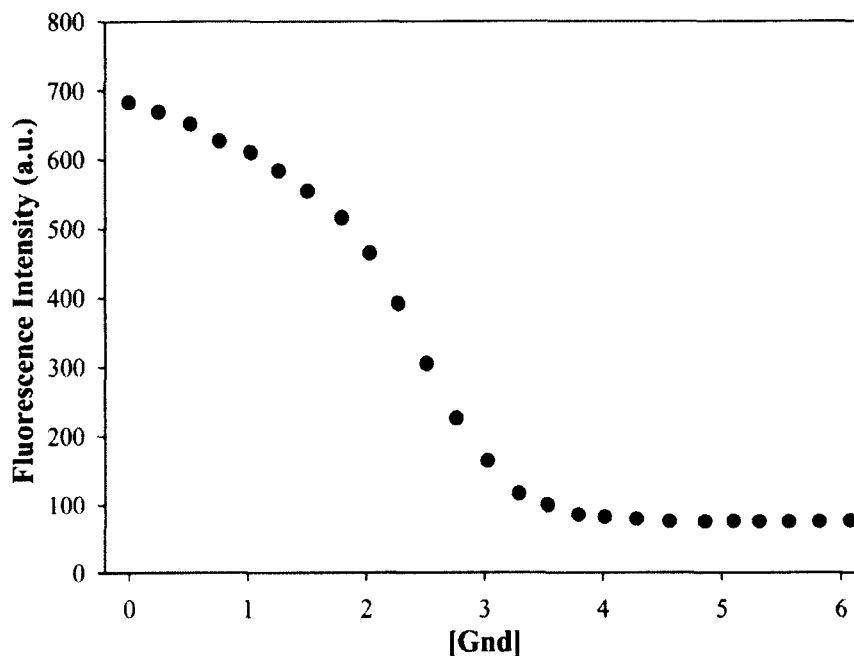


Figure 54. Equilibrium folding monitored by intrinsic tryptophan fluorescence.

Fluorescence data at collected at $\lambda_{em} = 330$ nm for WT-GB1 unfolding is plotted against Gnd-HCl concentration. We can see the slanted baseline for the native state between 0-1 M Gnd-HCl and a flat baseline for the unfolded state between 4-6 M Gnd-HCl. The data was plotted and visualized in SigmaPlot (Ver. 12.5)

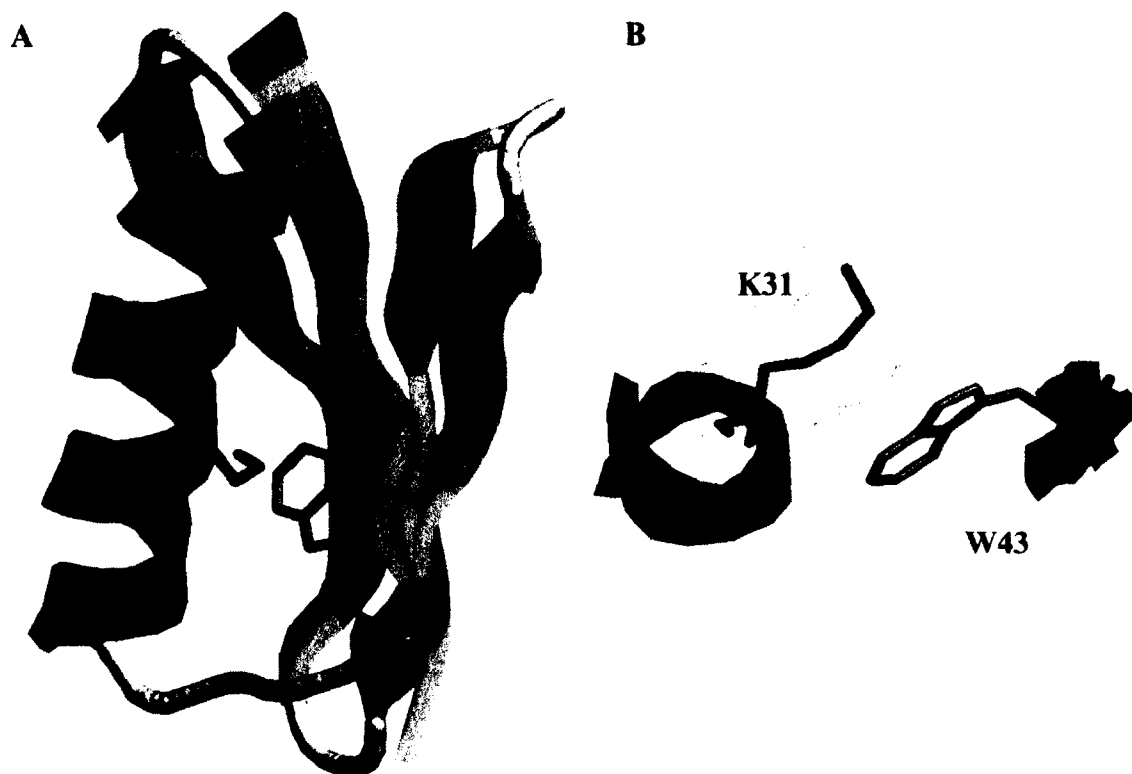


Figure 55. Image of the Lys31 and Trp43 proposed cation- π interaction. Secondary structure α -helix and β -strand are indicated by the colors pink and yellow respectively. The interaction was visualized in the whole protein (A) and a close-up cross-section (B). The residues were visualized with RasMol (Ver. 2.7.2.1.1).

The fractional unfolding of the WT and Phe52Tyr variant was calculated and plotted against Gnd-HCl concentration (Figure 56). In both proteins the folding is reversible. The midpoint of the transition occurs at ~2.57 M and ~2.68 M Gnd-HCl for the WT unfolding and refolding curves respectively. For the Phe52Tyr variant the midpoint transition occurs at ~1.90 M and ~1.99 M Gnd-HCl for the unfolding and refolding curves, respectively. Visual inspection of the curves as well as the ~0.68 M shift in the transition midpoint indicates that there is a loss in stability of the Phe52Tyr

variant relative to the WT. The calculated ΔG for WT-GB1 is -4.46 kcal/mol and there is a 0.66 kcal/mol loss in Gibbs free energy for Phe52Tyr to -3.80 kcal/mol. However, although there is a small decrease in stability it appears that the introduction of the tyrosine does not change the native structure (Figure 52).

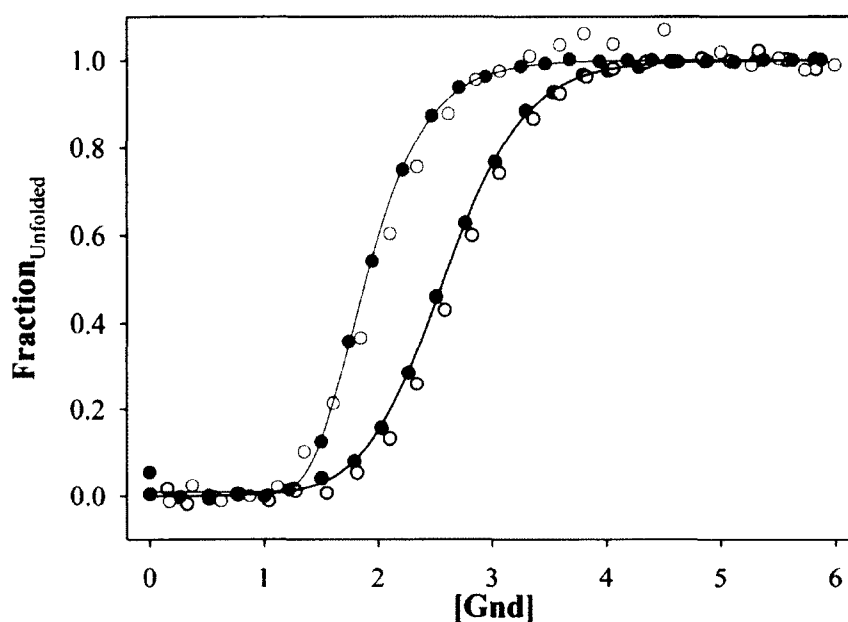


Figure 56. Fraction of unfolded protein versus Gnd-HCl. The fraction of the unfolded population is plotted against Gnd-HCl concentration for both WT (black circles) and Phe52Tyr (red circles) variant. Unfolding (filled circles) and refolding (unfilled circles) experiments were conducted to show reversibility of the folding for both proteins. Lines are sigmoidal regression fits for the unfolding curves. The data is plotted and fitted in SigmaPlot (Ver. 12.5)

Since there is a loss in stability in the equilibrium folding studies we wanted to see if the real-time folding kinetics changes significantly. Rapid-mixing stopped-flow studies were conducted to determine the folding rates for the WT and Phe52Tyr variant. From the equilibrium investigation conditions were selected to ensure the protein is in a completely unfolded state and transitions into conditions conducive to forming the native state fold. We selected 4 M Gnd-HCl as the unfolded conditions and induced refolding by a 7:1 dilution into refolding buffer, resulting in a 0.57 M final concentration which is well into the native state. Refolding of 1 mg/ml was done at 10 °C into 40% glycerol refolding buffer for both WT and the Phe52Tyr variant (Figure 57). Proteins were refolded at a flow rate of 6 ml/s, with a dead time of 6.8 ms. Due to the fast rate of folding and the large dead time, the unfolded baselines could not be shown. However, the unfolded baseline for WT was at ~0.57 and the Phe52Tyr variant unfolded baseline was at ~0.63. The folding kinetics for both proteins was fit to a double exponential regression and both the WT and Phe52Tyr variant fold rapidly in a biphasic pathway. The WT protein's first folding phase, with an amplitude of 95 % occurs at a rate of 145.39 s^{-1} (relaxation time was 6.88 ms). The second phase has an amplitude of 5 % with a folding rate of 9.94 s^{-1} (relaxation time was 100.60 ms). Similarly for the Phe52Tyr variant, 95 % of the initial folding collapse and organization occurred in the first phase with a folding rate of 153.57 s^{-1} (relaxation time was 6.51 ms). The second phase is much slower with the remaining 5 % of the folding occurring at a rate of 21.06 s^{-1} (relaxation time was 47.48 ms). The difference in folding kinetics between the WT and Phe52Tyr variant occurs in the second phase of folding, in which Phe52Tyr folds in about half the time required for the WT (Table 9). This indicates that although the initial collapse of the protein occurs similarly

and most of the native structure has formed the tyrosine mutation has an impact on the final stages of the “fine-tuning” phase in which the native structure is potentially stabilized by non-productive interactions which shows the folding process due to the need for rearrangement.

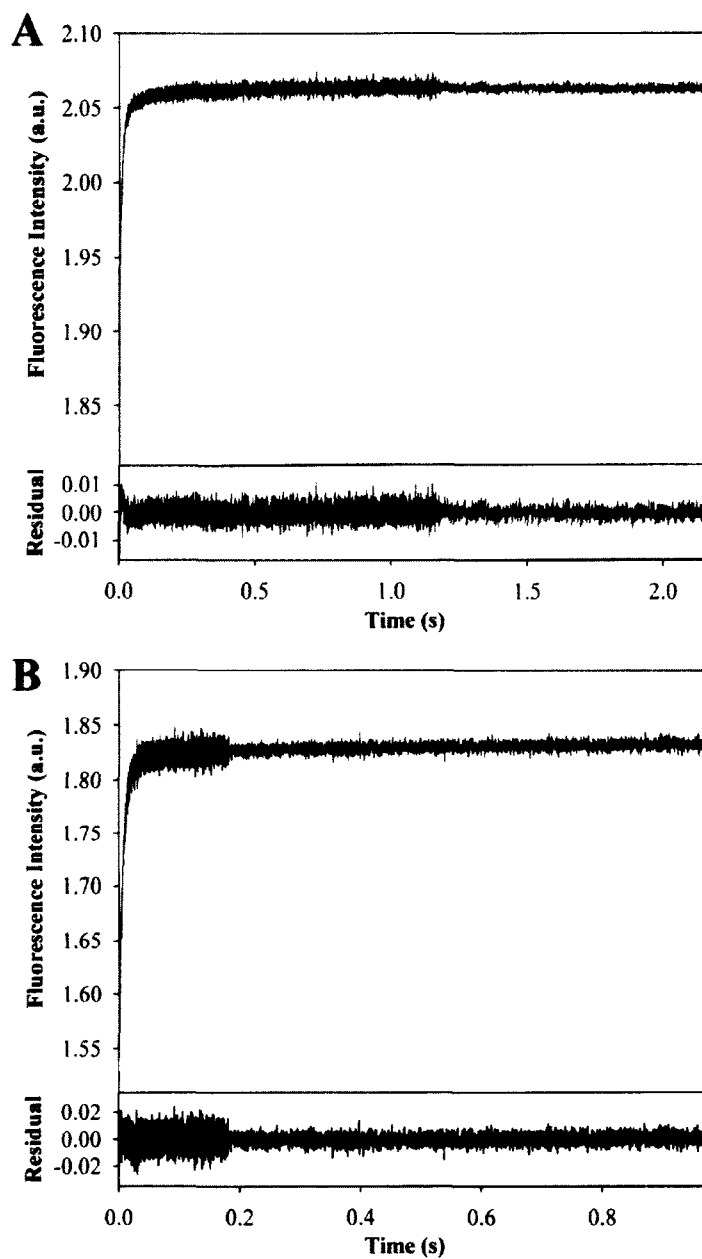


Figure 57. Stopped-flow kinetics of WT- and Phe52Tyr-GB1. Refolding kinetics of (A) WT- and (B) Phe52Tyr-GB1 with the resulting residuals. Curves were plotted and fit with a double exponential equation in SigmaPlot (Ver. 12.5). The red line indicates the double exponential regression line. The residuals of the fit are shown.

Table 9. Summary of stopped-flow kinetics data for WT- and F52Y-GB1

Protein (GB1)	Amplitude		Rate (<i>k</i>)		Relaxation Time	
	I	II	I	II	I	II
WT	95 %	5 %	145.39 ±0.951 s ⁻¹	9.94 ±0.223 s ⁻¹	6.88 ms	100.60 ms
F52Y	95 %	5 %	153.57 ±0.778 s ⁻¹	21.06 ±0.263 s ⁻¹	6.51 ms	47.48 ms

MATERIALS AND METHODS

All reagents were analytical grade. Tris base, NaCl, yeast extract, sodium azide (NaN_3) were purchased from VWR. (Gibbstown, NJ). Carbenicillin disodium salt, sodium dodecyl sulfate (SDS) running buffer and tryptone medium were obtained from Teknova (Hollister, CA). G-75 Sephadex medium size and Q-Sepharose Fast Flow anion exchange resins were from GE Healthcare (Pittsburgh, PA). Isopropyl- β -D-thiogalactoside (IPTG) was purchased from IBI Scientific (Peosta, IA). NUPAGE Novex 12% Bis-Tris Gel was obtained from Invitrogen (Grand Island, NJ). Competent *Escherichia coli* BL21(DE3) cells were purchased from Agilent Technologies Stratagene (Santa Clara, CA). Yeast extract was purchased from Research Organic (Cleveland, OH). Buffers were prepared and filtered through a 0.45 μm filter purchased from Pall Life Sciences (Suwanee, GA). Molecular porous membrane (1 kDa MWCO) for dialysis was obtained from Spectrum Laboratories (Rancho Dominguez, CA).

Structural Alignment and Percent Identity

To construct a structural alignment for GB1 the DaliLite v. 3 Sever was employed to determine proteins with similar structure (http://ekhidna.biocenter.helsinki.fi/dali_server/start). The DaliLite server is a comprehensive search method that surveys the protein data bank and does a sum-of-pairs comparison of superimposable structures. This method produces a measure of similarity by comparing intramolecular distances and calculating a similarity measure called the Dali-Z score. Structures that are significantly similar have a Z-score above 2, and usually have similar folds [434-436, 448]. The sequence identity of the structural alignment was

obtained from DaliLite structural data. Sequence alignments of selected structures needed to be significantly divergent with a sequence identity less than 25% and a broad range of functional diversity so that similarities obtained would be related to structure. Structural modifications were made by hand based on the visual comparison of the side-chain orientation in each selected structure using RasMol (Ver. 2.7.2.1.1). This manual analysis is required to ensure that the obtained structural alignment is aligned properly as computer algorithms are imperfect.

Conservation Analysis and Hydropathy

Once a structural alignment is completed and the side chain orientation is verified we can calculate position specific conservation. The number of each residue type at each position in the superfamily sequence alignment is calculated by summation of each type of amino acid. A program written for us by the He group (ODU Computer Science Department) was used to calculate these values. With the number of amino acids of each type at each position, entropy can be calculated by the following equation:

$$S(i) = - \sum_{j=1}^m \{P_j(i) \ln[P_j(i)]\} (i=1,2,3,\dots,14)$$

In the equation $P_j(i)$ is the fractional occurrence of each amino acid type j at each residue position i and m is the number of amino acids types possible in the particular analysis [449]. Since our structural sequence alignment incorporated fourteen sequences, i ranges from 1 to 14. Positional entropy tells us about the amino acid variability at each position. High entropy indicates high variability and thus infers low conservation and vice versa. Thus, to calculate conservation the following equation from [402] is used:

$$C(i) = \frac{1-S(i)}{\ln(m)}$$

From the analysis, residue positions whose conservation exceeds 0.45 are considered highly conserved whereas conservation between 0.45 and 0.30 are considered moderately conserved. Any conservation values lower than 0.30 are considered to be less conserved. Positions containing more than one gap are considered less conserved and are given a value of zero and considered non-conserved. To calculate residue specific hydrophobicity as it relates to the structural superfamily, the average hydrophobicity of all the amino acids at the selected position are summed according to the following equation:

$$\text{Hydrophobicity} = \sum_{j=1}^m j(i)H_j(i)$$

In the equation, $j(i)$ is the amino acid type j at residue position i in each structural alignment and $H_j(i)$ is the hydrophobicity of the amino acid. The summation goes from amino acid type $j=1$ to m which is the maximum number of amino acid types possible. The hydrophobicity values used were adapted from the most commonly used amino acid hydrophobicity [450]. The data from both conservation and hydrophobicity analysis were analyzed and plotted using SigmaPlot (Ver. 12.5, Systat Software)

Primer Design and Site-directed Mutagenesis

WT-GB1 cDNA was kindly provided by Professor Angela Gronenborn (Member of the National Academy of Science) from Pennsylvania State University in a pET-11a vector (Figure A4). The vector contains an antibiotic resistance gene that expresses the carbenicillin resistant β -lactamase allowing for specific selection of bacterium containing the cDNA. The GB1 cDNA was amplified using NEB 5 α competent *E. coli* (New

England Biolabs) and expressed in 50 ml of Luria Broth (LB) with carbenicillin (200 µg/ml). Amplified plasmid cDNA was extracted and purified using a Strataprep plasmid miniprep kit (Stratagene). Wild-type GB1 cDNA was mutated to Phe52Tyr GB1 using QuikChange II Site-Directed Mutagenesis (Agilent Technologies, Santa Clara, CA) (IBC biosafety protocol #12-006). The 50 µl PCR mutagenesis reactions consisted of the following components: 5 µl of 10X reaction buffer, 1 µl of the required 100 ng/µl cDNA template of WT-GB1, 1 µl of both forward and reverse primers, 1 µl dNTP mix and 1 µl of the provided 2.5 U/µl *pfuTurbo* DNA polymerase. The reaction was incubated in 16 cycles according to the following parameters per cycle: 95 °C for 30 sec (denaturation), 55 °C for 1 min (annealing) and 68 °C for 6 min (elongation). Primers were designed using the Agilent primer design program (<http://www.genomics.agilent.com/primerDesignProgram.jsp>) for all the selected conserved residues and for the control less conserved residues (Table 10). The primer for the mutation Phe52Tyr, sense primer 5'-CGACGCTACCAAAACCTACACGGTAACCGAATAGG-3', and antisense primer 3'-GCTGCGATGGTTTTGGATIGTGCCATTGGCTTATCC-5' were used to mutate the GB1 cDNA. Stability and amplification of the resulting mutant cDNA was verified and performed by expression in NEB 5α competent *E. coli* and purified using Strataprep plasmid miniprep kit. The resulting purified Phe52Tyr plasmid was stored at -20 °C until used. The insertion of the point mutation was verified by sequencing analysis at the Medical College of Virginia - Virginia Commonwealth University Nucleic Acids Research Facilities (Richmond, VA).

Table 10. Primers designed for mutational analysis of GB1

Codon Mutation	5'-3' Primer	3'-5' Primer
Y3A	TAAGAAGGAGATATACATATGCA <u>GGC</u> CAAGCTTATCCTGAACGGTA AAAC	ATTCTTCCTCTATATGTATACGTC <u>CGG</u> TCGAATAGGACTTGCCATT TG
L5A	AGATATACATATGCAGTACAAG <u>G</u> <u>C</u> TATCCTGAACGGTAAACCCTG	TCTATATGTATACGTCATGTT <u>CCG</u> ATAGGACTTGCCATTTTGGGAC
I6A	AGGAGATATACATATGCAGTACA AGCTT <u>GCC</u> CTGAACGGTAAACC C	TCCTCTATATGTATACGTCATGTT CGAA <u>CGG</u> ACTTGCCATTTTGGG
T16G	CGGTAAACCCTGAAAGGTGAAG <u>G</u> CACCACCGAAGC	GCCATTTTGGGACTTTCCTACTT <u>CC</u> GTGGTGGCTTCG
T18G	GAAAGGTGAAACCACCG <u>G</u> GCGAA GCTGTCGACGCT	CTTTCCTACTTTGGTGG <u>CCG</u> CTTCG ACAGCTGCGA
F30A	TGCTACCGCGGAAAAAGTT <u>GCC</u> A AACAGTACGCTAACGAC	ACGATGGCGCCTTTTCAAC <u>CGG</u> TT TGTCATGCGATTGCTG
Y33A	GCGGAAAAAGTTTTCAAACAG <u>G</u> C CGCTAACGACAACGGTGTTG	CGCCTTTTCAAAAGTTTGT <u>CCG</u> G CGATTGCTGTTGCCACAAC
Y45A	GTTGACGGTGAATGGACCG <u>G</u> CGA CGACGCTACCAAAAC	CAACTGCCACTTACCTGG <u>CGG</u> CT GCTGCGATGGTTTTG
D46A	GGTGAATGGACCTACG <u>C</u> GACGC TACCAAAACC	CCACTTACCTGGATGCG <u>G</u> GCTGCG ATGGTTTTGG
F52Y	CGACGCTACCAAAACCT <u>A</u> CACGG TAACCGAATAGG	GCTGCGATGGTTTTGGAT <u>T</u> GTGCCA TTGGCTTATCC
V54A	TACCAAAACCTTCACGG <u>C</u> AACCG AATAGGATCCGG	ATGGTTTTGGAAGTGCC <u>G</u> TTGGCT TATCCTAGGCC
T55G	CTACCAAAACCTTCACGGTAG <u>G</u> C GAATAGGATCCGGC	GATGGTTTTGGAAGTGCCAT <u>CCG</u> CTTATCCTAGGCCG

*Nucleotides changed from the WT-GB1 codons are indicated as bold and underlined

Protein Expression and Purification

The expression and purification protocol was adapted from previously published work (IBC #12-006) [386]. The protocol for WT- and Phe52Tyr-GB1 expression was optimized by multiple expressions with for example, increasing IPTG over variable time

to assess conditions for maximum protein expression. The WT and Phe52Tyr cDNA were then transformed into BL21 (DE3) competent *E. coli* (New England Biolabs) and grown in 6 L of LB with carbenicillin (200 µg/ml) until an optical density of 0.6-0.8 (Abs₆₀₀) was observed which indicates the mid-log phase of the bacterial growth. Protein expression was induced using 0.4 mM IPTG for 4 hours at 37 °C and 200 rpm agitation. Bacterium were harvested and sonicated for 1-3 hours on ice in 20 mM tris base (pH 8.5). The bacterial lysate was heated at 80 °C for 10-15 minutes then centrifuged in a Rotanta 460R Hettich centrifuge at 11500 rpm for 30 minutes to pellet cellular debris and aggregated proteins. The supernatant contained predominantly GB1 protein and was further purified using anion-exchange column chromatography. GB1 was eluted using a 0-500 mM NaCl sodium gradient in 20 mM tris base (pH 8.5) buffer over 480 min. Peaks containing GB1 were dialyzed in a 1000 MWCO membrane in double deionized H₂O and lyophilized in a Labconco freeze dryer (Kansas City, MO) Beach, VA). Lyophilized protein was further purified using G-75 sephadex (medium grain) size-exclusion column chromatography in 50mM tris base, 200 mM NaCl and 0.005 % NaN₃ (pH 7.5). GB1 protein purity was verified by SDS-PAGE and mass spectrometry and final purified protein was similarly dialyzed and dried. Protein was stored at -20 °C until use.

SDS-PAGE was run on NuPAGE Novex 12% Bis-Tris mini gels at 150 V for 1 hour. 15-20 µl of each sample and marker was loaded on to the gel. The isolated gel was stained with Coomassie brilliant blue stain for a minimum of 3 hours and then destained overnight. Staining and destaining solutions were made according to the protocols from Invitrogen.

Denaturation Equilibrium Fluorescence

Protein was taken from -20 °C storage, weighed and dissolved in 0.1 M tris base (pH 7.0) to produce stock protein solutions. Concentration was determined by measuring absorbance at 280 nm and dividing by the calculated molar extinction coefficient of 1.42 ml·mg⁻¹·cm⁻¹ [451]. Protein was diluted to a working concentration of 0.05 mg/ml for all equilibrium unfolding and refolding experiments. Protein was denatured by 0.25 M step-wise increases in Gnd-HCl concentration per sample up to a 6 M Gnd-HCl maximum. Samples were analyzed in triplicate on a Cary Eclipse fluorescence spectrophotometer (Varian, Palo Alto, CA) by excitation at 295 nm and emissions spectra were collected from 315 to 415 nm. The slit widths were 5 and 10 nm for excitation and emissions, respectively. Each spectrum was the average of 10 scans at 20 °C. Blanks were measured similarly and subtracted from the initial spectra to produce final fluorescence data. Denaturation curves were obtained by plotting emissions at 330 or 350 nm against Gnd-HCl concentration. The denaturant concentration of each sample was determined by measuring the refractive index using a hand-held ATAGO refractometer.

To determine the stability (ΔG) was calculated from the equilibrium data by determining the equation to the curves for the native and unfolding baselines as well as the transition region [12]. Using the following equations the fraction unfolded (f_U) is determined:

$$f_U = \frac{y - y_N}{y_U - y_N}$$

In the equation, y is the observed fluorescence for each concentration of Gnd-HCl. The calculated equation of the line for the baselines corresponds to y_N and y_U for

native and unfolded respectively. The equilibrium constant K_{eq} at each concentration of Gnd-HCl is calculated according to the following equation:

$$K_{eq} = \frac{f_U}{1 - f_U}$$

The ΔG is dependent on the denaturant concentration and is calculated for each of the equilibrium constants of the transition region using the following equation:

$$\Delta G_{[Gnd]} = -RT \ln(K_{eq})$$

R is the gas constant ($1.986 \text{ cal} \cdot \text{K}^{-1} \cdot \text{mol}^{-1}$) and T is the temperature of the experiment. To obtain the ΔG^0 independent of the denaturant the $\Delta G_{[Gnd]}$ at each position in the transition region is plotted versus Gnd-HCl concentration. A linear regression fit of this plot allows for the determination of the ΔG^0 according to the following equation where m is the slope of the line:

$$\Delta G_{[Gnd]} = \Delta G^0 - m[Gnd]$$

Circular Dichroism Analysis

WT- and Phe52Tyr-GB1 were analyzed by far- and near- UV circular dichroism in 0.1 M tris base (pH 7.0). A stock solution of each protein was diluted in buffer to a final working concentration of 0.2 mg/ml for far-UV and 0.5 mg/ml for near-UV spectropolarimetry. Samples were measured with a Jasco J-815 spectropolarimeter using continuous scan mode set to a rate of 200 nm/min. Each experiment is run in triplicate and is the average of 10-15 scans. The mean residue ellipticity (Θ_{MRW}) is calculated using the following equation:

$$\Theta_{MRW} = \frac{MRW \cdot \Theta_{obs}}{10 \cdot d \cdot c}$$

In the above equation, $MRW = M/(N - 1)$, where M is the molecular mass of the polypeptide chain (in Da), and N is the number of amino acids in the chain. Θ_{obs} is the observed ellipticity (degrees), d is the pathlength (cm), and c is the protein concentration (g/ml).

Stopped-Flow Folding Kinetics

Folding kinetics studies were conducted using a SFM-400 stopped-flow system (Bio-Logic, France). WT-GB1 and Phe52Tyr variant protein were denatured in 0.1 M Tris (pH 7.0) and 4 M Gnd-HCl for a minimum of 3 hours or overnight. Refolding was initiated by a seven-fold dilution into refolding buffer containing 0.1 M Tris-HCl and 40 % glycerol (pH 7.0) refolding buffer at 10 °C with a flow rate of 6 ml/s. Kinetics experiments were run using a 1.5 mm FC-15 cuvette and a mixing dead time of 8.6 ms. Fluorescence changes were monitored by excitation at 295 nm and emissions detection >320 nm using a bandpass filter (Semrock, Rochester, NY). The slit widths were both 1 mm for excitation and emissions for all experiments. Spectra were obtained from the average of 5-10 repeat experiments. Curves were analyzed and fitted to an exponential regression using SigmaPlot (Ver. 12.5, Systat Software).

CHAPTER IV

DEVELOPMENT OF A METHOD TO ELUCIDATE THE FORMATION OF A LONG-RANGE INTERACTION IN THE B1 DOMAIN OF PROTEIN G USING NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY

OVERVIEW

Protein folding is a multifaceted problem in biochemistry and has been studied using high-resolution for years to try and unravel the code that allows for rapid formation of a protein native structure in a fraction of a second. Proteins fold along an energy landscape by forming specific non-covalent short- and long-range interactions in an ordered process which results in the proper organization of structural components into a native conformation. These short- and long-range interactions are guided by intermolecular forces such as hydrogen bonding, electrostatic interactions, hydrophobicity and disulfide bonds. Long-range interactions are classified as interactions between amino acids that are greater than 7 residues from each other in the primary structure but within 5 Å in the tertiary structure. Whereas, short-range interactions are less than 7 residues and are typically found in secondary structure elements [360, 404, 452]. These interactions can be monitored by biophysical analysis and indirect methods. Understanding the underlying structural interactions of proteins is important in identifying key determinants that dictate the mechanism of protein folding and misfolding. We have previously discussed that the Trp43 is partially buried in the

hydrophobic core and is sensitive to conformational changes in the folding of GB1 (Table A1) (Refer to Chapter III). The mechanism of protein folding can be expressed as an increase in long-range interactions that arrange the overall topology while short-range interactions order the backbone structures such as α -helices, β -hairpins and β -turns [453]. We must be able to understand how and why specific long-range interactions form first and how this process is facilitated by amino acid character and the aqueous environment. Using NMR spectroscopy it is possible to visualize the long-range interactions that make up the tertiary structure of a protein using both solution-state and magic-angle spinning solid-state NMR (MAS-NMR) [454-461].

Studying protein folding has typically been done with indirect biophysical methods such as fluorescence, absorbance and circular dichroism [462]. A high resolution method such as quenched-flow hydrogen-deuterium exchange coupled to NMR has become more specialized for protein folding but is only ideal for short-range interactions such as hydrogen bonds in the α -helices and β -sheets [405]. However, atomic resolution studies involving the formation of specific long-range interactions in real-time during the folding process are virtually absent in the literature. In this chapter we intend to discuss a novel method of folding-freezing stopped-flow spectroscopy in combination with MAS-NMR to follow the interaction of three specific amino acids in the folding of GB1 as a 'proof-of-concept' method. The ^{13}C - ^{13}C CRPK's between the selected amino acids are proposed to be specifically visualized using MAS-NMR. In summary, we propose an approach and preliminary data to monitor the formation of long-range interactions between Phe 30, Trp 43 and Phe 52, three amino acids found in the core of the protein. Changes in these three amino acids will also be investigated as a function of

pH using HSQC NMR. Preliminary solution-state NMR studies are also conducted to facilitate characterizing the protein structure under different environmental conditions.

To reiterate, GB1 is a small 56-residue protein that functions as an immunoglobulin binding domain of the cell surface bacterial protein G. This protein was selected as a model system because of the ease of expression, no disulfide bond, lack of free cysteine, ability to fold at high concentrations, solved NMR and crystal structures in addition to fundamental investigations of the folding of GB1 make it an ideal candidate for this method. With this selected model system several advances were made to develop its applicability. Efforts to establish conditions to synthesize site-specifically ^{13}C -labelled protein and monitor the formation of a conserved set of long-range interactions in the core of GB1 using equilibrium unfolding in concert with high-resolution NMR is an important first step. To the best of my knowledge the formation of specific conserved long-range interactions has never been shown with atomic resolution. We selected residues 30, 43 and 52 for ^{13}C -labeling as all three residues are conserved, found in the core of GB1 and play a role in several long-range interactions between them. We have already established reversible equilibrium unfolding conditions for WT-GB1 (Refer to Chapter III). We will also show the kinetics of folding of GB1 at high concentrations required for NMR which is another critical step. The results from chapter III enabled the following work described in this chapter to be conducted. Lastly, preliminary NMR studies were conducted to assess appropriate conditions and parameters.

RESULTS AND DISCUSSION

Synthesis of the ^{13}C -Phe/ ^{13}C -Trp labelled protein was done using an engineered *E. coli* auxotroph (Aux). This strain is unable to synthesize all three aromatic amino acids due to a knockout in the chorismate synthase gene in the Shikimate pathway (Figure 58). Chorismate is a precursor molecule for phenylalanine, tryptophan and tyrosine. Carefully cultured Aux bacteria were λ DE3 lysogenized, incorporating the gene for T7 RNA polymerase into the genome of the Aux (Aux(DE3)). Incorporation of the T7 RNA polymerase gene allows for the overexpression of plasmid genes regulated by the *lac* operon and which selectively have a T7 promoter (Figure 59). The synthesis of the T7 polymerase was verified using a T7 tester phage which is a phage mutant that has a T7 deletion and is completely defective unless the host cell provides T7 RNA polymerase. Host cells successfully lysogenized will provide the T7 RNA polymerase and cause the production of plaques or halos in the presence of IPTG (Figure 60). On the other hand, significantly smaller plaques are observed in the absence of T7 RNA polymerase. Cultured Aux(DE3) was verified by a titration of tester phage in the presence of IPTG. We also were able to determine how tightly regulated the T7 RNA polymerase is in the newly synthesized Aux(DE3). The growth of Aux(DE3) without IPTG and the size of the resulting plaques indicated basal levels of T7 RNA polymerase expression. We needed to ensure that there was no leaky expression of T7 RNA polymerase to enable efficient expression of ^{13}C -labelled protein using IPTG.

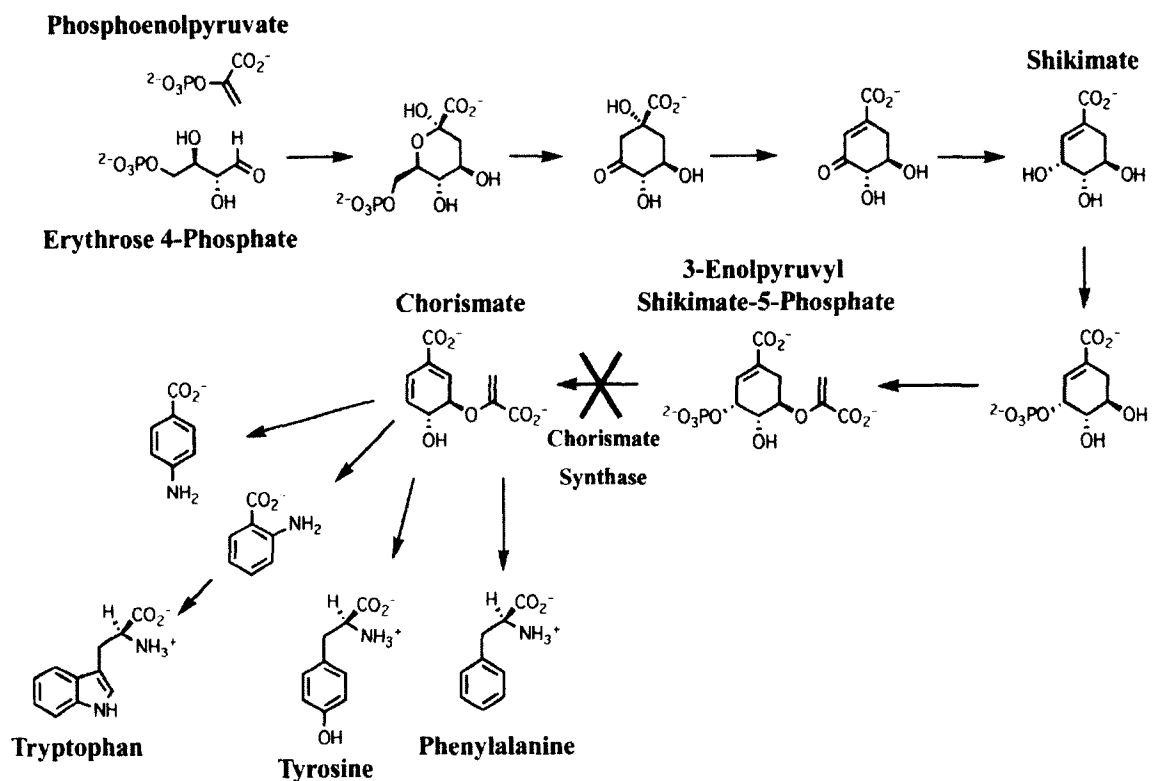


Figure 58. Schematic of the Shikimate pathway showing the synthesis of aromatic amino acids in *E. coli*. Our Aux *E. coli* bacteria have a knockout in the enzyme that synthesizes chorismate the primary precursor to aromatic amino acids. Figure adapted from [463].

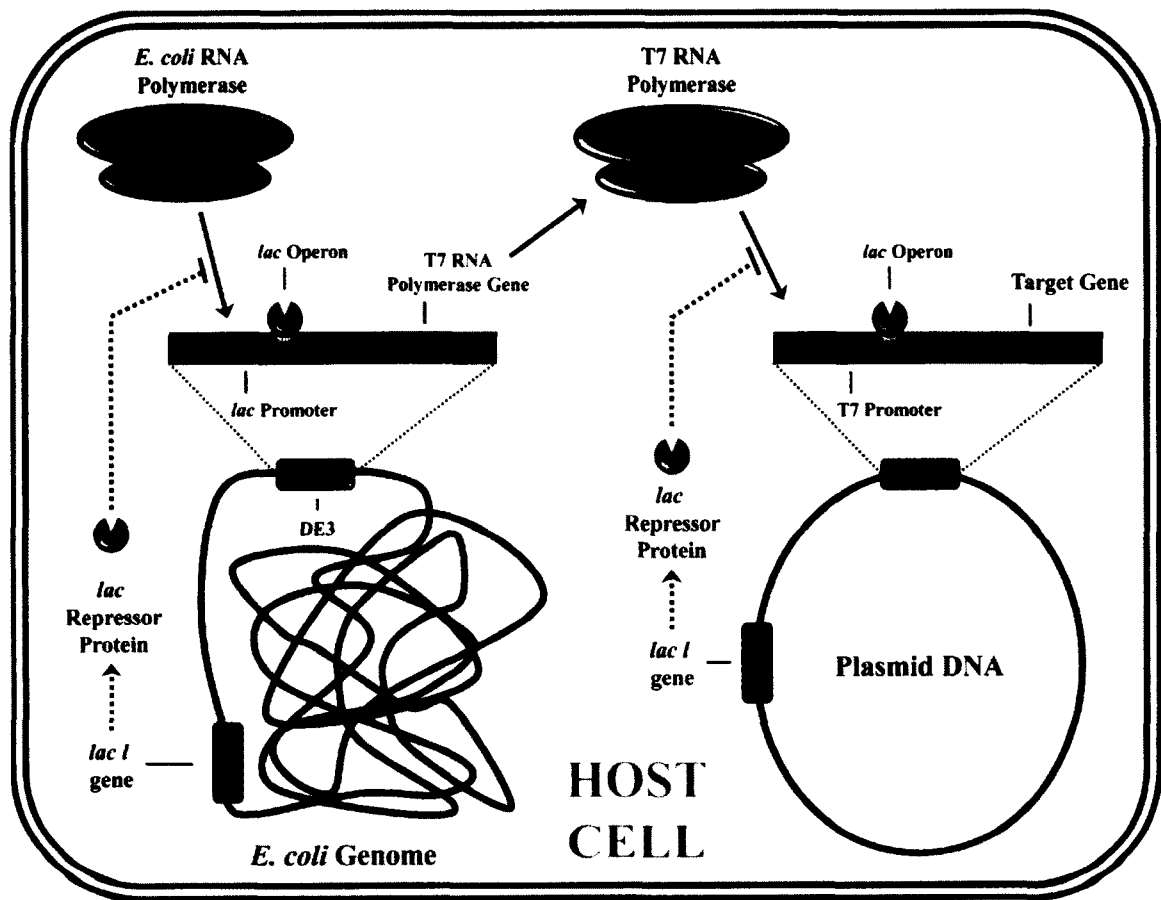


Figure 59. Schematic of the Lac Operon, T7 Promoter and its regulation of recombinant gene expression in *E. coli*. Induction by IPTG results in the dissociation of the repressor protein bound to the lac operon and enables the polymerase to bind to its corresponding promoter [464, 465]. Figure was drawn from information obtained in [466].



Figure 60. Titration of tester phage with auxotrophic lysogens. (A) Undiluted tester phage. (B) 20X dilution. (C) 50X dilution. (D) 2,500X dilution. (E) 125,000X dilution. (F) Undiluted tester phage control (No IPTG).

To ensure that the Aux(DE3) is a true auxotroph we tested the ability of the cells to grow in the absence of each aromatic amino acid (Figure 61A). If the Aux(DE3) is a true auxotroph there should be little to no growth of the cells in the absence of Trp, Phe or Tyr independently. It is clear from Figure 61A that there is no growth of the cells in comparison to the control containing all three aromatics. The cultures were grown for ~24 hours and we wanted to determine if the cells were still viable so a small amount was streaked onto fresh LB plates to see if they would grow (Figure 61B). The cells were able

to recover and grew quite well. Now that the Aux(DE3) was verified we could now use them to selectively label the WT-GB1 protein.



Figure 61. Aromatic knockout verification of cultures of auxotrophic *E. coli* transfected with Phe52Tyr-GB1. (A) Minimal media cultures; Aromatics (+), Tyr(-), Phe(-) and Trp(-) from left to right. 100 µl of (B) Tyr(-), (C) Phe(-) and (D) Trp(-) cultures plated after ~24 hours of incubation.

To investigate the folding and formation of a specific set of non-covalent long-range interactions we synthesized a uniquely labeled protein using the previously discussed Aux(DE3). We initially intended to investigate only a single set of interactions between Phe30 and Trp43 found in the Phe52Tyr-GB1 variant. However, biophysical analysis of the stability of Phe52Tyr-GB1, previously discussed in chapter III, indicated to us that the instability of Phe52Tyr-GB1 could cause problems at the high concentrations required for NMR (Refer to Chapter III). Thus, we selected the WT-GB1 for specific and uniform ^{13}C -labeling. The interactions we are interested in are between Phe30-Trp43 and Phe52-Trp43 (Figure 62A). Interestingly, Phe30, Trp43 and Phe52 are highly connected in the structure of GB1 having a significant number of long-range interactions which indicates their importance (Figure A5). We used ^{13}C -Phe/ ^{13}C -Trp labeled ^{13}C -Phe and ^{13}C -Trp for the specific long-range interactions investigated (Figure 62B). For these interactions to be sensitive to NMR spectroscopy, they need to be within 5 Å distance from one another. The Phe30- βC is 4.24 and 4.30 Å from the Trp while the Phe52- βC is 4.42 and 4.56 Å from the Trp ring (Figure 63). These interactions are within the van der Waals radius and find each other in the core of WT-GB1 and all three amino acids are conserved as previously discussed (Chapter III). It is interesting to note that typically protein expression in minimal media result in lower yields than traditional rich LB media however in the present GB1 expression there was little difference in expression levels.

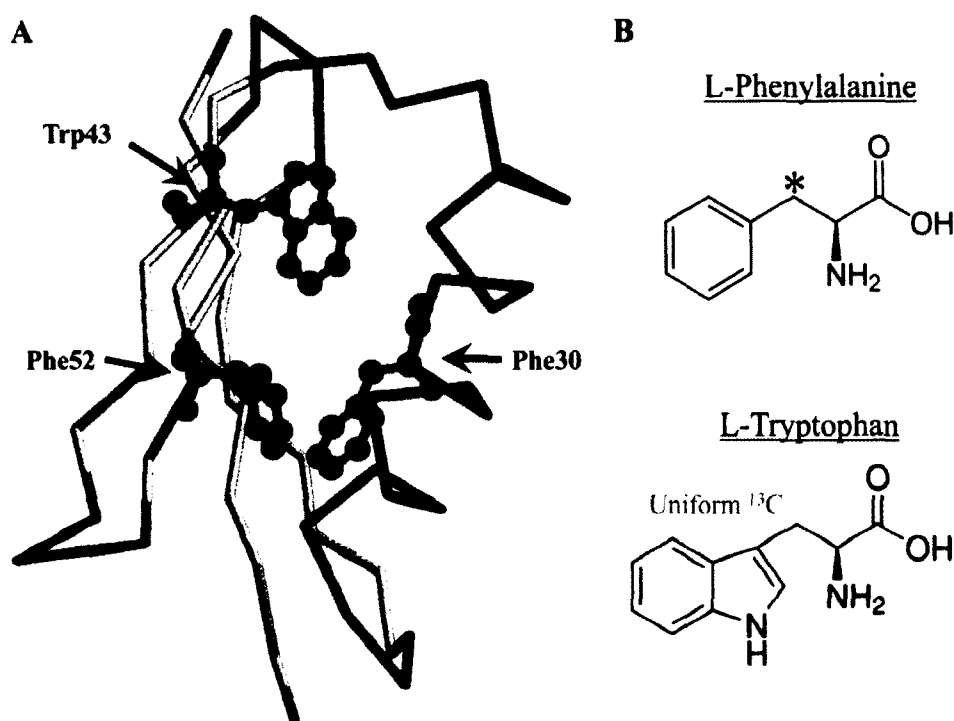


Figure 62. Backbone structure of GB1 with highlighted core aromatics and specific ^{13}C labeling of Phe and Trp. (A) Structure of the WT-GB1 backbone. Yellow and magenta portions indicate secondary structures β -strands and α -helix, respectively. Amino acids Trp (cyan) and Phe (red) are shown as ball and stick models. (B) Molecular structure of Phe and Trp, and their respective ^{13}C labeling positions in red. The residues are visualized with RasMol (Ver. 2.7.2.1.1).

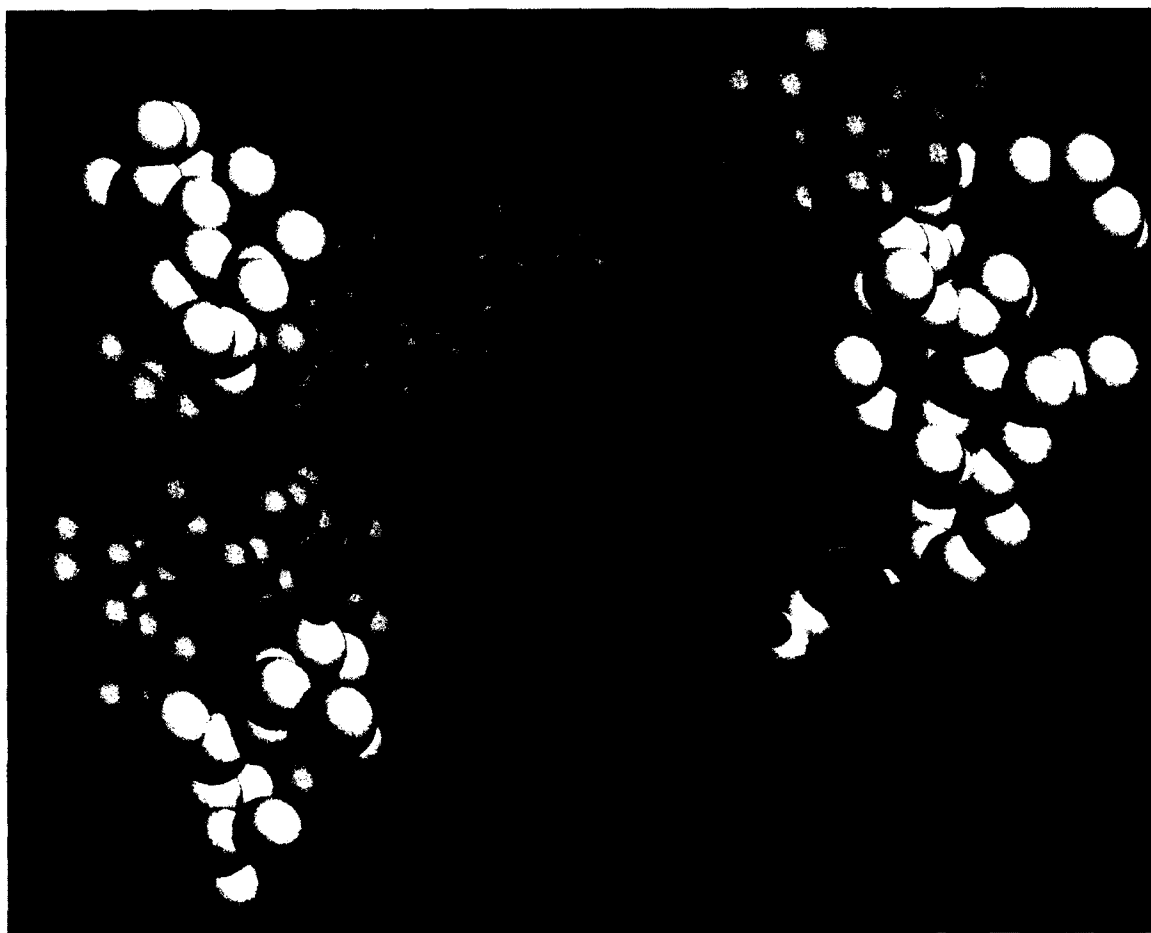


Figure 63. ^{13}C -Labeled carbons of interest in the core structure of GB1. Yellow, magenta and white portions indicate regions of β -, α - and loop structures, respectively. Amino acids Phe30, Trp43 and Phe52 are shown as wireframe (green). The measured distances for the interactions between Phe30 and Trp43 (dashed cyan lines) and Phe52 and Trp43 (solid cyan lines) are shown. This structure was visualized with RasMol (Ver. 2.7.2.1.1).

The purpose of using the Aux to synthesize selectively labeled protein is in preparation for ssNMR in which the specific labeling lends itself to our folding method because NMR on uniformly isotopically labeled proteins would produce an abundance of carbon signals. Therefore, the resonance peaks would be indistinguishable and identifying and monitoring specific interactions would be quite difficult (Figure A6). However, with very specific labelling of the Trp43 and two Phe- β C's we can theoretically map the interactions as they come together or separate in space (Figure 64). In the example schematic below, we can see that as the carbons separate in space due to the transition from folded to unfolded there is a loss in CRPK signal (Figure 64). In the end we want to be able to map all the long-range interactions (Table A2) in the folding of GB1 using this Aux(DE3) system of expression coupled with folding kinetics and NMR spectroscopy.

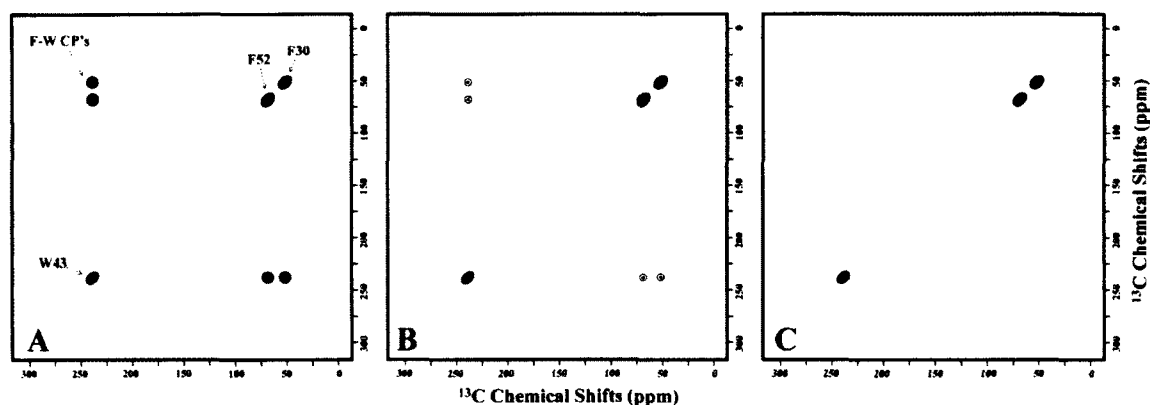
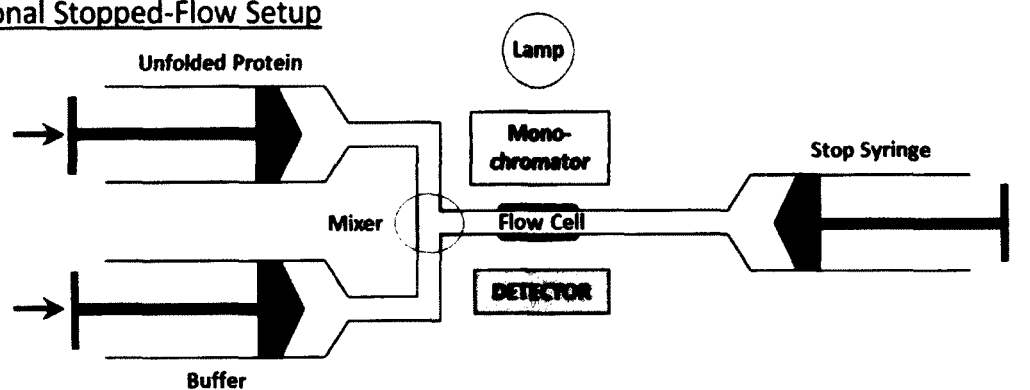


Figure 64. Representative schematic of anticipated cross-peak formation from interactions of ^{13}C -specifically labeled WT-GB1. Blue peaks indicated Phe30 and Phe52 while the red peak indicates the Trp43. Purple peaks indicate Trp-Phe CRPK's. The folding transitions from (A) folded to (B) intermediate to (C) unfolded states.

The next stage of this project was to capture a long-range interaction in the core of WT-GB1 using a novel folding-freezing method we are developing. An attachment to the stopped-flow system has been designed and engineered to potentially capture a long-range interaction by instantly freezing a protein sample in supercooled isopentane (-80 °C) during the folding process (Figure 65). The design of the freezing folding stopped-flow experiment requires a denaturation process and a carefully characterized protein folding system. The method we intended to use for the unfolding and refolding transition is chemical denaturation by Gnd-HCl. GB1 works well in this aspect as the folding transition is reversible by Gnd-HCl denaturation as previously discussed (Refer to Chapter III). However, this method proved to have limitations in its application and coupling with NMR spectroscopy. To work with this method the required amount of protein needed in order to be observable by MAS-NMR may be outside the range of possibility for protein folding. We needed to see how much protein we could conceivably pack into a 4 mm solid-state rotor using the chemical denaturation method. Because chemical denaturation requires a significant dilution in order to initiate refolding, we also wanted to know if it was possible to fold WT-GB1 at exceedingly high concentrations due to its robust structural stability. We were able to dissolve WT-GB1 at concentrations of up to 120 mg/ml in 4 M Gnd-HCl without any apparent precipitation. We investigated the folding kinetics of these high concentrations of WT-GB1 and Phe52Tyr-GB1 at initial concentrations of ~100 mg/ml with a 7-fold dilution down to a final concentration of ~15 mg/ml (Figure 66). The folding kinetics showed that in both WT- and Phe52Tyr-GB1 there was still a biphasic folding transition however there was a shift in the folding distribution. At low concentrations the distribution was 95 % and 5 % for the first and

second phase, respectively, however at higher concentrations a larger portion of the folding event occurred over the second phase with a distribution of ~70 and 30 % for both WT- and Phe52Tyr-GB1. For WT-GB1 at high concentrations the folding rate decreases to 100.34 s^{-1} (9.97 ms relaxation time) and 17.25 s^{-1} (57.99 ms relaxation time) for the first and second phases, respectively. For Phe52Tyr-GB1 the folding rate decreases to 119.82 s^{-1} (8.35 ms relaxation time) and 26.21 s^{-1} (38.16 ms relaxation time) for the first and second phases. In comparison initial folding of the protein at high concentrations in both cases appeared slower followed by a more rapid second phase. This high concentration dependent slowing effect could be due to interactions of the unfolded polypeptides with one another that reduce the proteins ability to rapidly form native interactions or an increase in non-specific/non-native interactions within the protein. Interestingly, refolding the proteins at a higher dilution factor (~13-fold dilution) resulted in a slight change in the refolding rates particularly for Phe52Tyr-GB1 (Table 11). In summary, the biphasic distribution between high and low concentrations with respect to amplitudes and to some degree the rates of the second phase differ (Refer to Chapter III). This finding suggests there are no serious complications to using high protein concentrations in refolding kinetics.

Traditional Stopped-Flow Setup



Freezing-Folding Stopped-Flow Setup

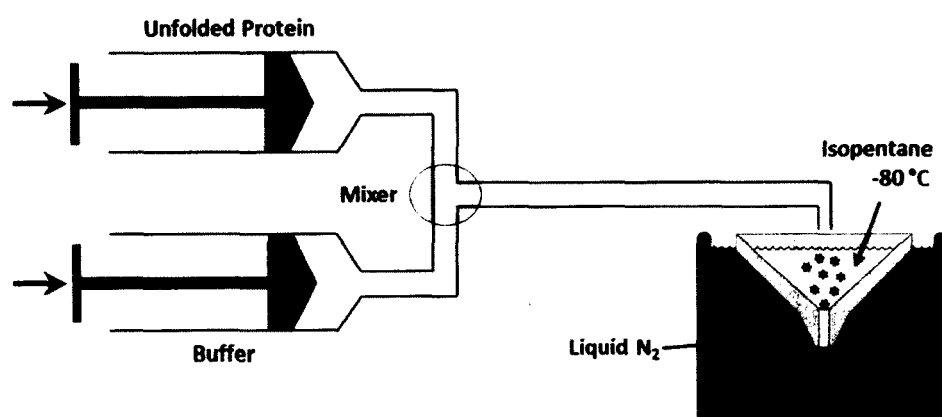


Figure 65. Schematic outline of a traditional versus folding-freezing stopped-flow instrument. In a traditional stopped-flow instrument the denatured protein in one syringe is rapidly mixed with refolding buffer from another syringe in the follow cell and monitored. The detector could be, for example, fluorescence, absorption or CD. In the modified folding-freezing apparatus the solutions are mixed similarly, however detection is removed and the protein reaction is sprayed into supercooled isopentane. The resulting crystals can be packed into a solid-state rotor for MAS-NMR. This Figure was redrawn and adapted from [153].

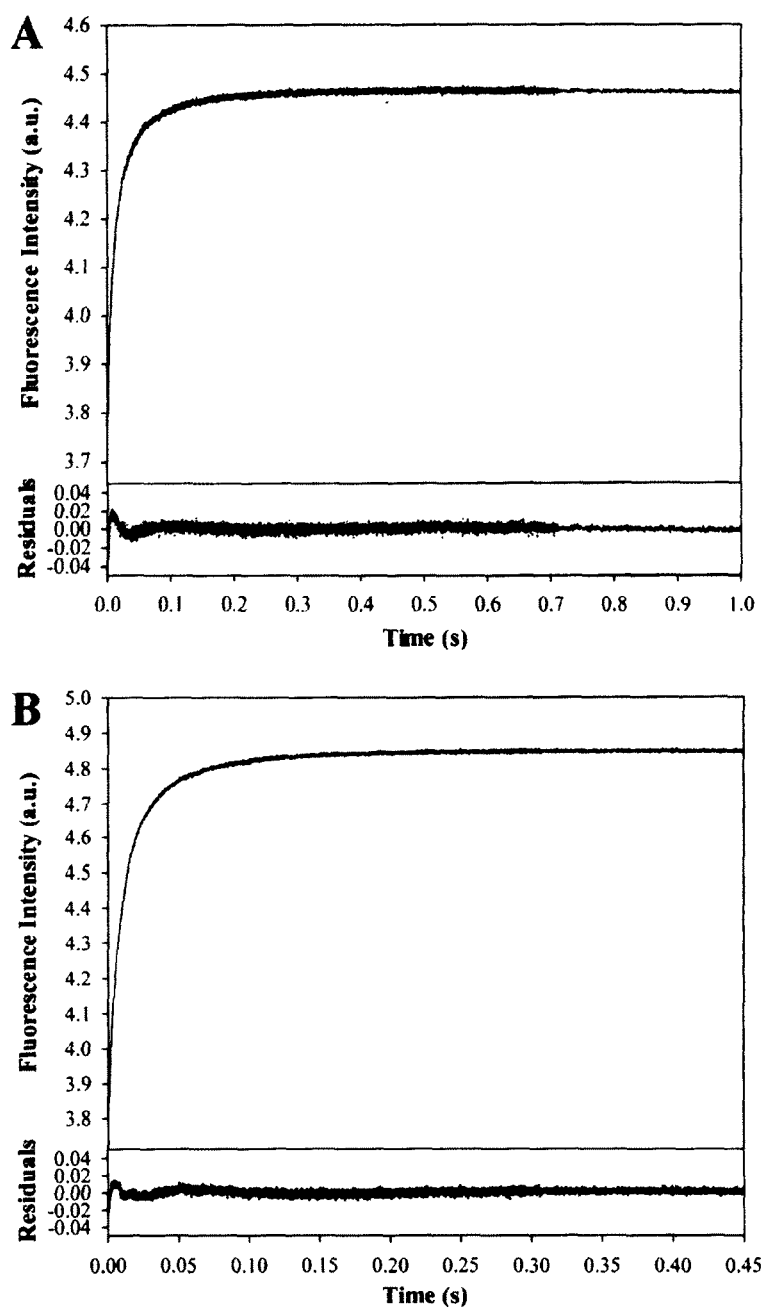


Figure 66. Stopped-flow fluorescence kinetics of WT- and Phe52Tyr-GB1 at high protein concentrations. Refolding kinetics of (A) WT- and (B) Phe52Tyr-GB1 with the resulting residuals. Curves were plotted and fit with a double exponential equation in SigmaPlot (Ver. 12.5). The red line indicates the double exponential regression line.

Table 11. Stopped-flow kinetics data of WT- and Phe52Tyr-GB1 at high concentrations

Protein (GB1)	Dilution	Amplitude		Rate (<i>k</i>)		Relaxation Time	
		I	II	I	II	I	II
WT	7X	70 %	30 %	100.34 $\pm 0.316 \text{ s}^{-1}$	17.25 $\pm 0.079 \text{ s}^{-1}$	9.97 ms	57.99 ms
WT	13X	66 %	34 %	106.88 $\pm 0.514 \text{ s}^{-1}$	17.12 $\pm 0.113 \text{ s}^{-1}$	9.36 ms	58.40 ms
F52Y	7X	72 %	28 %	119.82 $\pm 0.230 \text{ s}^{-1}$	26.21 $\pm 0.071 \text{ s}^{-1}$	8.35 ms	38.16 ms
F52Y	13X	73 %	27 %	125.62 $\pm 0.340 \text{ s}^{-1}$	19.27 $\pm 0.062 \text{ s}^{-1}$	7.96 ms	51.88 ms

*Initial protein concentrations were ~100-110 mg/ml in all experiments

From the folding kinetics experiments we determined that we would be able to get ~1 mg of protein into the solid-state 4 mm rotor for our experiments. However this amount was insufficient in acquiring ssNMR signal because the WT-GB1 only has a ^{13}C mass percent of 2.5, which meant we needed much more protein. At this point we began to characterize the ^{13}C -Phe/ ^{13}C -Trp labeled WT-GB1 in order to explore the possibility of a pH unfolding refolding jump and determine concentrations required to get significant signal in ssNMR. In addition we considered a high temperature rapid freezing method in which a heated sample is rapidly frozen in -80°C isopentane which would allow us to capture a folding intermediate with the long-range interactions of interest [455].

We initially characterized the structure of the ^{13}C -Phe/ ^{13}C -Trp labeled WT-GB1 in a 1D ^{13}C experiment using solution-state NMR (Figure 67). From the spectra we were able to assign peaks to specific labeled carbons using ACD/SpecManager (Ver 9.15). The peaks are well resolved and all the carbons that we anticipated to be labeled are visible. It

seems that the carbons on the two Phe's are also distinguishable and are labeled as 7a and 7b. Analysis of lyophilized ^{13}C -Phe/ ^{13}C -Trp labeled WT-GB1 protein in solid-state NMR using cross polarization MAS (CP-MAS) for ^{13}C analysis showed a similar spectrum (Figure 68). However, in the CP-MAS the signal is significantly broadened due to anisotropic interactions. In comparison to the uniformly labeled WT-GB1 spectrum (Figure 69) which shows a significant number of peaks corresponding to the 274 carbons in WT-GB1 the signal is much more resolved. The peaks in the uniformly labeled sample are overlapping and merged into large wide peaks corresponding to peak overlap of carbons in very similar environments. In the spectrum of the ^{13}C -Phe/ ^{13}C -Trp labeled WT-GB1 we see better resolved peaks however, because of the differences in anisotropic averaging the sensitivity and resolution is decreased with low separation of peaks (Figure 68) in comparison to solution-state NMR (Figure 67).

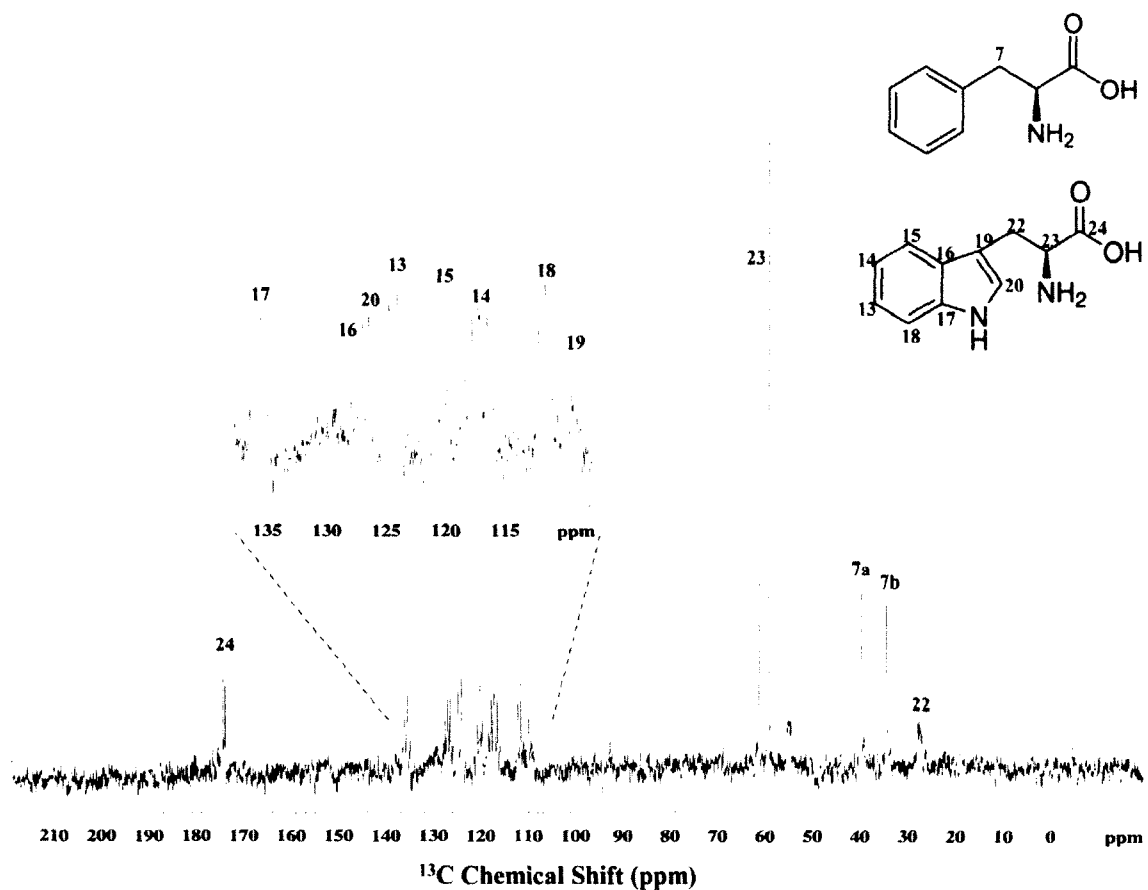


Figure 67. 1D ^{13}C solution-state NMR spectrum of ^{13}C -Phe/ ^{13}C -Trp labeled WT-GB1. The ^{13}C NMR spectrum was acquired on a 400 MHz Bruker Avance III with 1024 scans. The 10 mg/ml sample was analyzed in pH 7.0. The peak numbers correlate to carbon assignments for Trp43, Phe30 and Phe52.

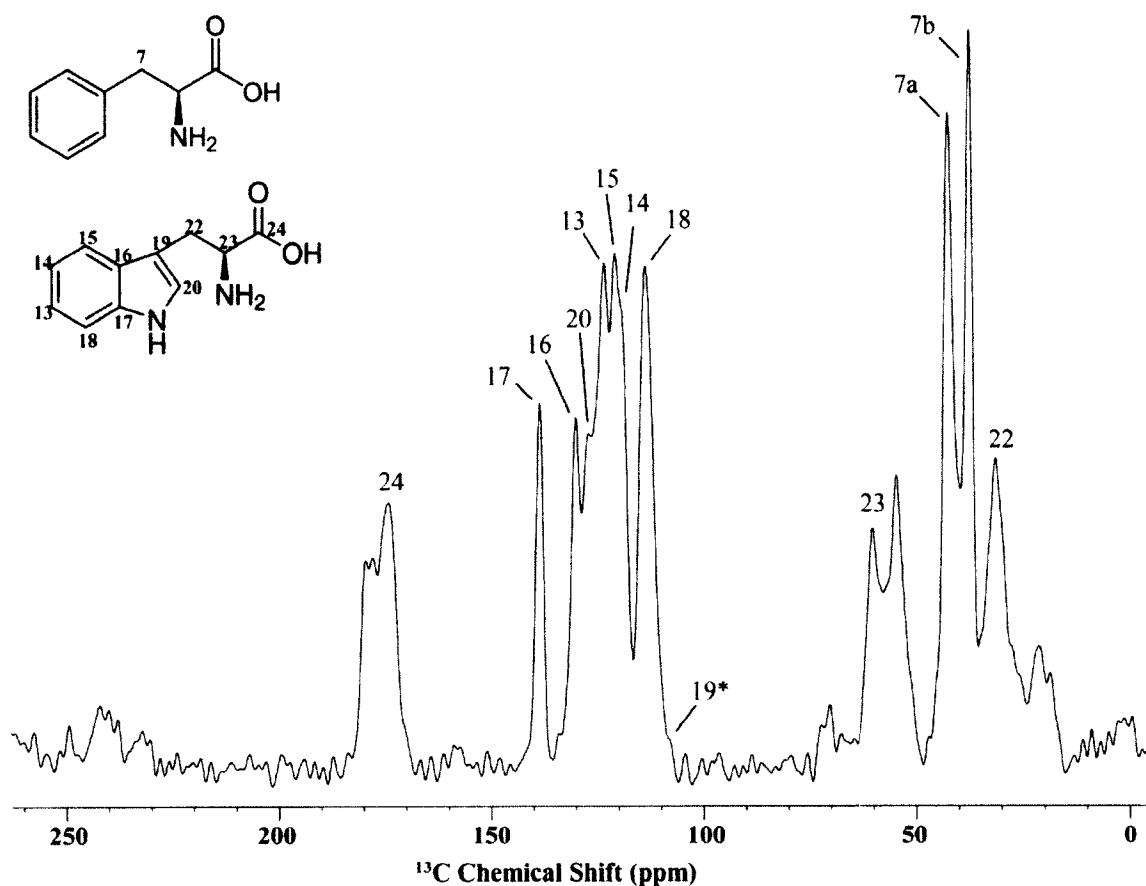


Figure 68. 1D ^{13}C ssNMR CP-MAS spectrum of ^{13}C -Phe/ ^{13}C -Trp labeled WT-GB1.

21 mg of lyophilized protein was packed into a 4 mm zirconia rotor and analyzed on a 400 MHz Bruker Avance II with 512 scans at a spin rate of 12 kHz were collected. The peak numbers correlate to carbon assignments for Trp43, Phe30 and Phe52. Position indicated by an * indicates anticipated peak not seen.

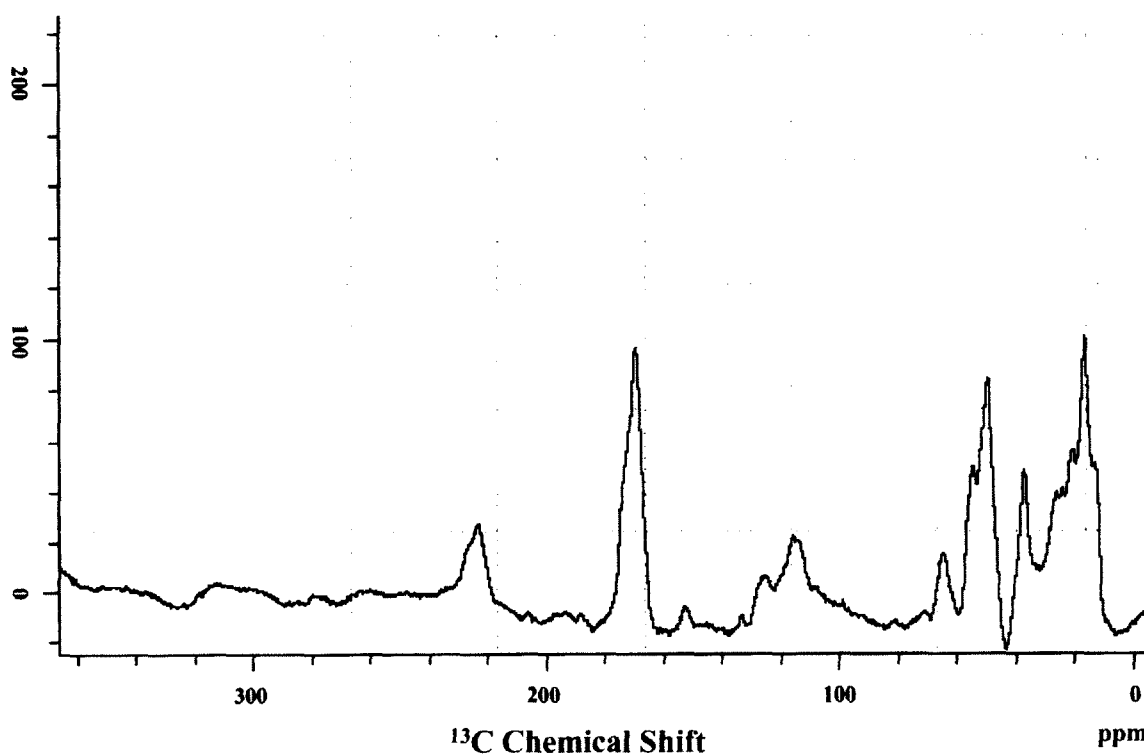


Figure 69. 1D ^{13}C ssNMR spectrum of uniformly labeled WT-GB1. 35 mg of lyophilized protein was packed into a 4 mm zirconia rotor and analyzed on a Bruker Avance 750 MHz NMR spectrometer. The sample was spun at 12 kHz during the experiment.

We initially characterized the structure of the uniformly labeled WT-GB1 in a 2D ^{13}C - ^{13}C DARR experiment (Figure 70). In the DARR experiment ^1H magnetization is transferred to ^{13}C -carbons to enhance the NMR signal followed by subsequent transfers of magnetization to other ^{13}C -carbons through space which appear as CRPK's in the 2D NMR spectra. The distance in magnetization transfer is dependent on the mixing time. In order for there to be a ^{13}C - ^{13}C transfer the two carbons must be close in space. In the uniformly labeled WT-GB1 experiment we see many CRPK's as all carbons are labeled.

This particular experiment sets the foundation for the specific interactions we intend to monitor as the protein is denatured. The DARR NMR spectrum showed many internal CRPK's between the labeled carbons of the Trp43 which was to be expected. More importantly, we see two CRPK's between the Phe and Trp at positions 144 and 47 ppm and 129 and 47 ppm. Both of these CRPK's are interactions of Trp43 benzene ring with the same Phe. However, based on the positions and interactions with the Trp43 benzene ring it's difficult to determine which Phe is interacting and which is not interacting. A possible explanation for one Phe not interacting with the Trp could be due to the fact that the sample is a dry powder and the intermolecular position of the second Phe may have transitioned to a position greater than 5 Å from the Trp. Running the sample in ssNMR as a frozen solution would provide greater understanding of differences between solid- and solution-state structural states. However, based on these results there was evidence that there was structural modification on the core structure as a result of lyophilization. In any case this result indicated that our ^{13}C -Phe/ ^{13}C -Trp labeled sample would provide ideal CRPK signal for at least one Phe and could be used to monitor this long-range interaction during the folding process. One potential problem with respect to protein concentration was that it required ~42 mg of solid protein analyzed for about 4 hours in order to acquire the clear present CRPK signal. Under solution conditions 42 mg translates into ~525 mg/ml which is extremely impractical. However, to overcome this we can reduce the amount of protein and increase the run time of the experiment by an order of magnitude (from ~4 hours to ~2 days). Also we could increase the rotor diameter from 4 mm to 6 mm, which would increase the volume from 80 μl to 180 μl and change the concentration to ~230 mg/ml. These quantities for folding maybe overestimated as the minimal amount

of protein frozen in solution remains to be determined and will be further discussed.

Ultimately however, it may be viable to refold with far less protein and this will also be discussed in future work. It is more amenable to use a combination of both to reduce the concentration to about 100 mg/ml. In addition, to decrease signal line broadening we can increase the magnetic field strength from 400 MHz to 750 MHz to increase the sensitivity and enhance the resulting spectrum [467]. Although it is theoretically possible it appears that chemical denaturation kinetics in concert with ssNMR detection seems to be on the edge of the detection limits of a 400 MHz ssNMR.

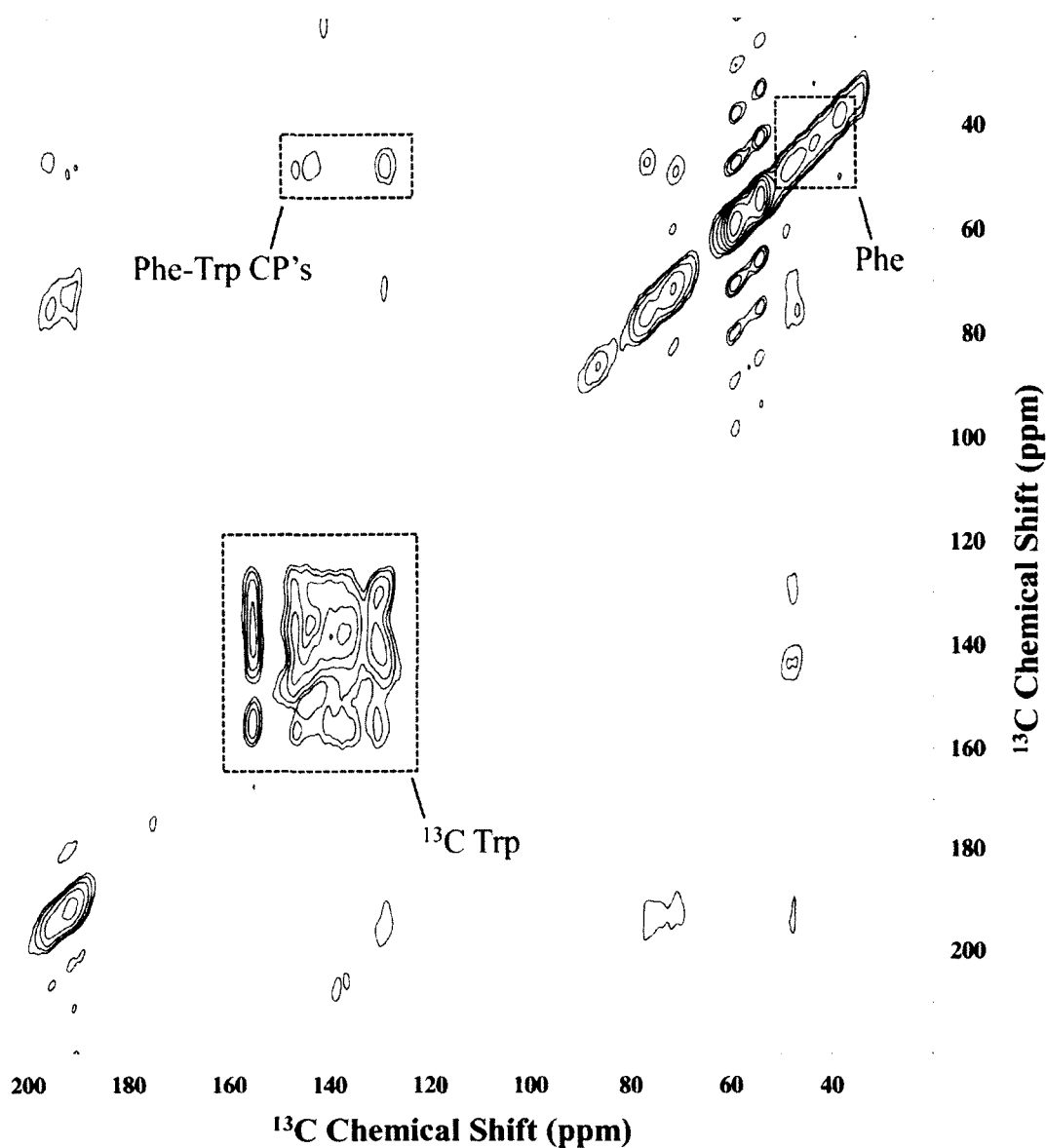


Figure 70. 2D ^{13}C - ^{13}C -DARR experiment on ^{13}C -Phe/ ^{13}C -Trp labeled WT-GB1. A lyophilized sample of 41.7 mg ^{13}C -Phe/ ^{13}C -Trp labeled WT-GB1 was packed into a 4 mm rotor and analyzed on a Bruker Avance II 400 MHz solid-state NMR spectrometer. The sample was spun at 12 kHz and 64 scans were acquired at ^{13}C frequency of 100.6 MHz.

We looked at the behavior of the ^{13}C -Phe/ ^{13}C -Trp labeled GB1 in a pH dependent manner using solution state NMR for two reasons. In the first instance we wanted to know about the proteins behavior at physiological versus extreme pH to better understand its robust nature. Second, we wanted to determine if WT-GB1 formed a typical partially unfolded molten globule state at low pH. This would potentially enable us to dissect and identify any core contacts as well as use a pH jump to monitor refolding of long-range interactions with potentially a smaller dilution than starting with a chemically denatured state. We conducted 1D ^1H (Figure 71) and 2D HSQC NMR (Figure 72) experiments at pH 2.0, 7.0 and 12.1. Peaks were assigned using ACD/SpecManager (Ver 9.15) to predict the 2D HSQC spectra of Phe and Trp with a ^1H , ^{13}C COSY correlation. In many cases the protein molten globule state can be induced *in vitro* by placing the protein in highly acid conditions [58, 59]. So in addition we looked at the changes that would indicate a transient molten-globule form. The 2D HSQC experiments showed that the environment around the labeled carbons did not change very much (Figure 73). This means that even under high and low pH environments the core of WT-GB1 is maintained as it relates to the environment around Phe 30, Phe52 and Trp43. It appears that at low pH GB1 is not induced into a molten-globule state because there are no shifts in the overall core environment. It's curious though we see an additional CRPK that I was unable to identify. Interestingly, prediction of a free Phe HSQC spectrum resulted in the 2 β -carbon hydrogens to be split by the neighboring proton identical to those peaks at the carbon shift of 37 ppm. Whereas if a partial backbone is included in the prediction this splitting pattern is no longer observed and a single peak at a carbon shift of 41 ppm is predicted. This could indicate that at pH 7 the two Phe residues are in orientations that are resulting

in two NMR chemical shift patterns (Figure 73). In one condition the environment is much more symmetrical resulting in a proton splitting pattern and in the other the environment the two protons are in asymmetrical conditions resulting in a single peak.

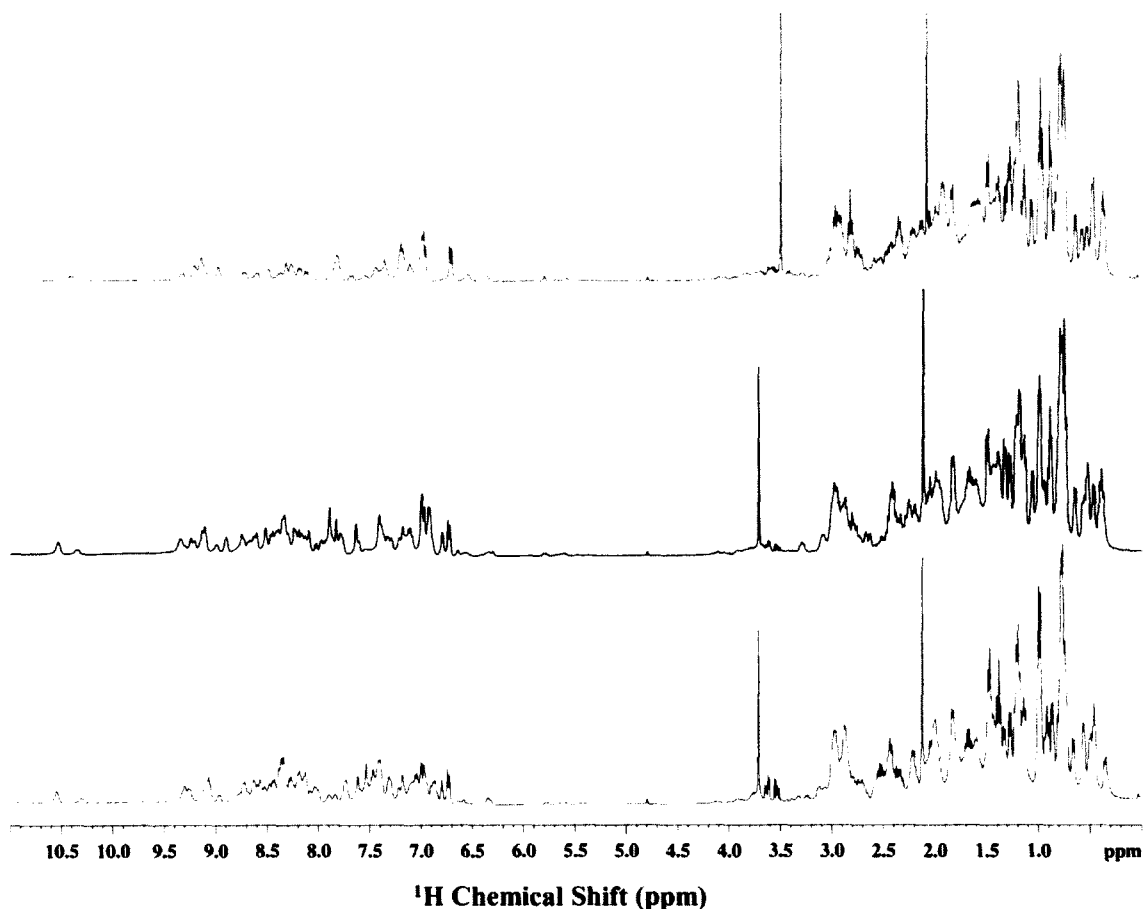


Figure 71. 1D ^1H NMR spectrum of ^{13}C -Phe/ ^{13}C -Trp labeled WT-GB1 versus pH. Specifically labeled WT-GB1 was analyzed in solvent at pH 2.0 (red), pH 7.0 (blue) and pH 12.1 (green). Samples were examined at 128 scans on a Bruker Avance II 400 MHz solutions NMR spectrometer with a ^1H frequency of 400.5 MHz.

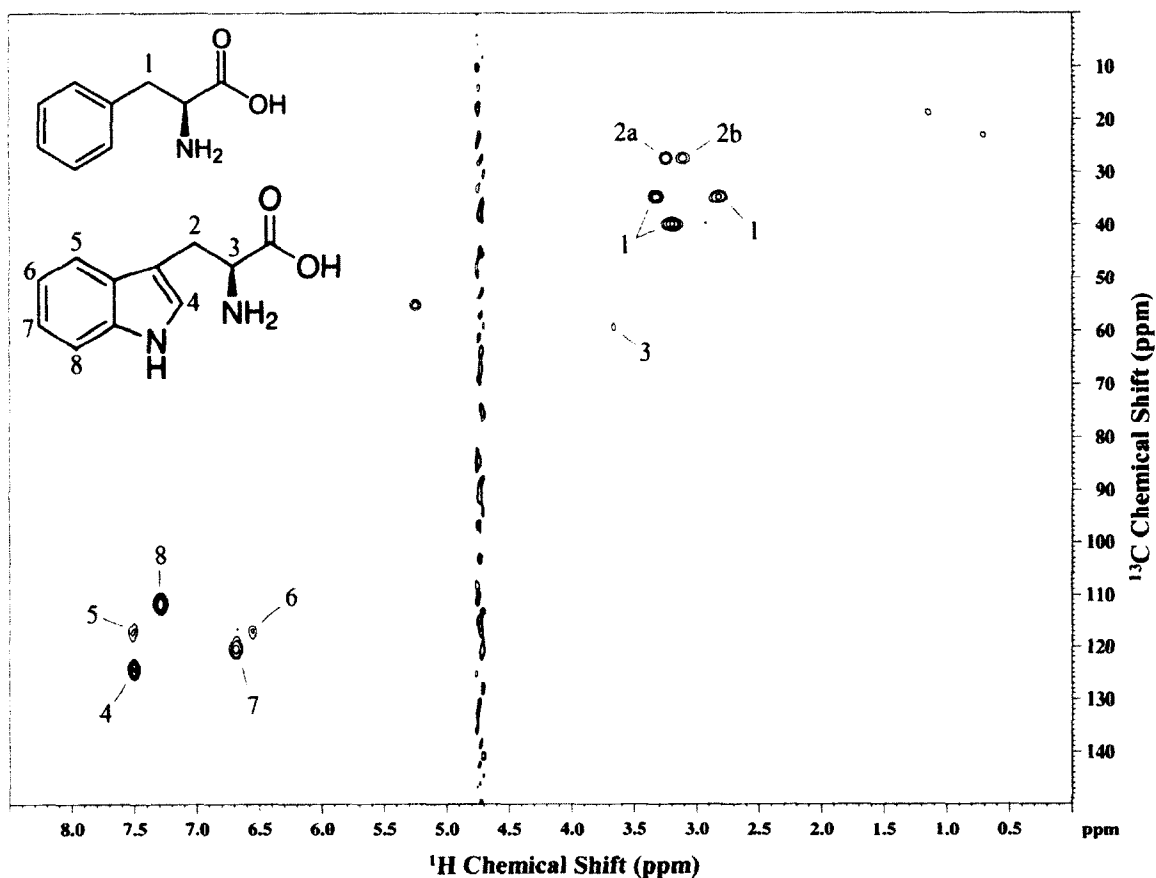


Figure 72. 2D ^1H - ^{13}C HSQC spectrum of ^{13}C -Phe/ ^{13}C -Trp labeled WT-GB1 with proposed chemical shift assignments. The HSQC spectrum was acquired on a 400 MHz Bruker Avance III with 64 scans. The 10 mg/ml sample was analyzed in pH 7.0 $\text{H}_2\text{O}/\text{D}_2\text{O}$. The peak numbers correlate to predicted ^1H - ^{13}C assignments for Trp43, Phe30 and Phe52. The sample was examined with H_2O presaturation. ^1H signal at 4.73 ppm is residual H_2O signal.

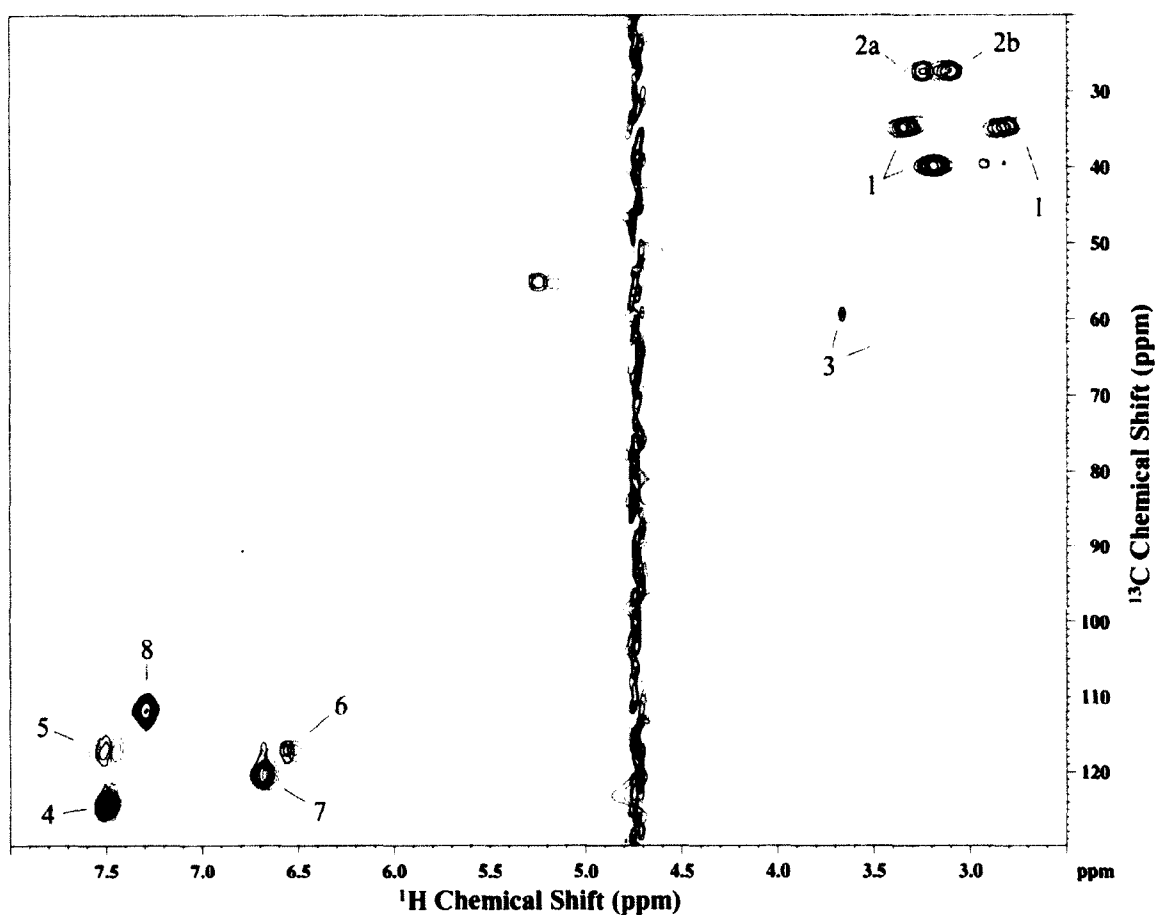


Figure 73. Analysis of pH dependence on the core of GB1 using ^1H - ^{13}C HSQC NMR.

Spectrum of ^{13}C -Phe/ ^{13}C -Trp labeled WT-GB1 at pH 2.0 (red), pH 7.0 (blue) and pH 12.1 (green). The HSQC spectrum was acquired on a 400 MHz Bruker Avance III with 64 scans. The 10 mg/ml sample was analyzed in 90 % H_2O /10 % D_2O solvent. The peak numbers correlate to ^1H - ^{13}C assignments for Trp43, Phe30 and Phe52 found in Figure 14. The sample was examined with H_2O presaturation. ^1H signal at 4.73 ppm is residual H_2O signal.

To get a better understanding of what kind of structural changes were occurring at pH 2.0, we further characterized WT-GB1 using CD (Figure 74). A typical molten globule state will have native-like secondary structure while having very little to no tertiary structure. Interestingly, at pH 2.0 WT-GB1 seems to have a significant decrease in secondary structure content while maintaining almost native tertiary structure. Interestingly, it may be possible that the protein has a gross native-like topology with minimal secondary structure, and hence be an ideal equilibrium transition-state like structure. The near-UV CD data matches well with the 1D ^1H and 2D HSQC solutions NMR experimental data observed at pH 2.0, in that there are only minor changes to the tertiary structure corresponding to minor changes to the core of GB1.

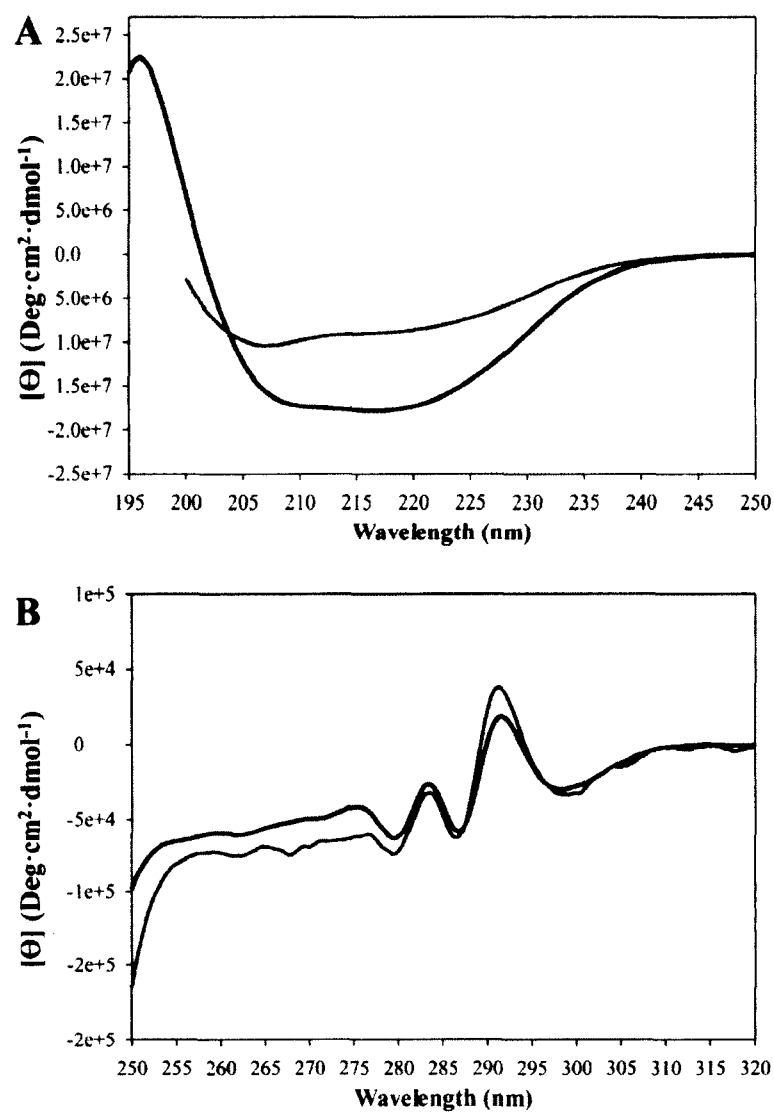


Figure 74. Molten globule analysis of WT-GB1 at pH 2.0 by CD. (A) Far-UV and (B) Near-UV CD of WT-GB1 at pH 7.0 (black) and pH 2.0 (green). The data was plotted and visualized in SigmaPlot (Ver. 12.5).

MATERIALS AND METHODS

Auxotrophic strain

An aromatic auxotrophic *E. coli* strain (K-12 derivative) with a knockout in the chorismate pathway to inhibit tryptophan, tyrosine and phenylalanine metabolism was obtained from the Coli Genetic Stock Center at Yale University and carefully cultured in LB media under sterile conditions. The chorismate synthase knockout was verified by culturing cells in minimal media supplemented with all the amino acids except for the selected amino acid targets: tryptophan, tyrosine and phenylalanine. Verified stock solutions of Aux cells were stored in 15 % glycerol at -80 °C until use.

To support the expression of genes regulated by the T7 promoter the Aux must be lysogenized to introduce the λ DE3 into the chromosome of the *E. coli*. Using a lysogenization kit the bacterium were grown in Luria broth supplemented with 0.2% maltose, 10 mM MgSO₄, and carbenicillin (200 μ g/ml) at 37 °C to an OD₆₀₀ of 0.5. Using stock lysates provided λ DE3 phage, helper phage, and selection phage were mixed with 1–10 μ l of the cells. Cells were incubated 37 °C for 20 min to allow phage to adsorb to host. The mixture was then plated onto an LB plate, covered inverted and incubate at 37°C overnight. Surviving colonies should be Aux(DE3). Cells were verified using a tester phage in the presence of IPTG for the presence of the λ DE3 insertion. Verified stock solutions of the selected colony Aux(DE3) were stored in 15 % glycerol at -80 °C until use.

Transformation, expression and purification of ¹³C labeled WT-GB1

Specifically labeled amino acids where obtained from Cambridge isotopes (Tewksbury, MA). WT-GB1 plasmid was transformed into the Aux(DE3) *E. coli* in the

Greene Lab (Old Dominion University), plated and cultured in minimal media containing the appropriate antibiotic and amino acid nutrients (IBC #12-006). Expression of uniformly ^{13}C -labeled WT-GB1 follows the previously established protocol discussed in chapter III except that the LB media is replaced with M9 minimal media and substituted with ^{13}C -glucose as the sole carbon source (1.5g/L). ^{13}C -Phe/ ^{13}C -Trp labeled WT-GB1 was expressed in M9 minimal media supplemented with each of the 20 amino acids. Table 12 outlines the quantities of each reagent in the minimal media for both expressions. ^{13}C -Phe was used to specifically label a single carbon at positions 30 and 52 that interact with the uniformly labeled ^{13}C -Trp at position 43. Purification of both the uniform and ^{13}C -Phe/ ^{13}C -Trp labeled GB1 followed the previously discussed protocol. Cells were lysed and purified using anion exchange and size exclusion chromatography to yield purified ^{13}C labeled Phe52Tyr-GB1. Minimal media from the ^{13}C -Phe/ ^{13}C -Trp labeled GB1 expression was reused in a subsequent expression to exhaust the ^{13}C -amino acids and maximize protein yield. The expression resulted in a significant amount of protein, 550 mg of the labeled ^{13}C -Phe/ ^{13}C -Trp protein and ~250 mg of uniformly labeled protein (IBC #12-006).

Table 12. Reagent list for ^{13}C -Phe/ ^{13}C -Trp labeled GB1 expression in Auxo(DE3)

Group	Reagent	Molarity (mM)
10X M9 Salts	Na_2HPO_4	33.7
	KH_2PO_4	22.0
	NaCl	8.6
	NH_4Cl	9.4
	CaCl_2	0.3
	MgSO_4	1.0
Mineral Solution	$\text{ZnSO}_4 \cdot 7\text{H}_2\text{O}$	0.003
	$\text{CuSO}_4 \cdot 5\text{H}_2\text{O}$	0.006
	$\text{CoCl}_2 \cdot 6\text{H}_2\text{O}$	0.005
	$\text{MnSO}_4 \cdot \text{H}_2\text{O}$	0.003
	H_3BO_3	0.004
Amino Acids	Ala	5.8
	Arg	2.0
	Gly	7.5
	His	0.8
	Ile	1.8
	Leu	1.8
	Lys	2.6
	Met	1.7
	Pro	0.9
	Ser	3.6
	Thr	2.0
	Val	2.1
	Asp*	3.3
	Asn*	3.1
	Cys*	0.4
	Gln*	2.8
	Glu*	4.5
	Tyr*	1.0
	^{13}C -Phe*	0.8
	^{13}C -Trp*	0.2
Additional Key Reagents	Carbenicillin*	0.9
	Vitamins*	1 %
	D-Glucose*	0.4 %
	4-amino benzoic acid*	0.001
	4-hydroxybenzoic acid*	0.001
	2,3-dihydroxybenzoic acid*	0.001

*Reagents sterilized via 0.2 μm syringe filter. All other reagents were sterilized by autoclave.

Stopped-Flow Folding Kinetics

Folding kinetics studies were conducted using a SFM-400 stopped-flow system (Bio-Logic, France). WT-GB1 and Phe52Tyr variant protein were denatured in 0.1 M Tris (pH 7.0) and 4 M Gnd-HCl for a minimum of 3 hours or overnight. Refolding was initiated by a 7- to 14-fold dilution into refolding buffer containing 0.1 M Tris and 40 % glycerol (pH 7.0) refolding at 10 °C with a flow rate of 6 ml/s. Kinetics experiments were run using a 1.5 mm FC-15 cuvette and a mixing dead time of 8.6 ms. Fluorescence changes were monitored by excitation at 295 nm and emissions detection >320 nm using a bandpass filter (Semrock, Rochester, NY). The slit widths were both 1 mm for excitation and emission for all experiments. Spectra were obtained from the average of 5-10 repeat experiments. Curves were analyzed and fitted to a double exponential regression using SigmaPlot (Ver. 12.5, Systat Software).

Circular Dichroism

WT-GB1 was analyzed by far- and near-UV CD in 0.1 M tris base (pH 7.0) and H₂O (pH 2.0) at 20 °C. A stock solution of each protein was diluted in buffer to a final working concentration of 0.2 mg/ml for far-UV and 0.5 mg/ml for near-UV spectropolarimetry. Samples were measured with a Jasco J-815 spectropolarimeter using continuous scan mode set to a rate of 200 nm/min. Each experiment is run in triplicate and is the average of 10-15 scans.

Nuclear Magnetic Resonance

Solution-state NMR samples were prepared in 90:10 H₂O/D₂O. Protein was weighed and dissolved in buffer to a final concentration of 10 mg/ml. 600 µl of sample was pipet into high quality NMR tubes for analysis. Samples were analyzed on an

Avance II 400 MHz NMR spectrometer (Bruker). For solid state NMR samples were weighed and carefully packed into a 4 mm MAS zirconia rotor. Experiments were conducted on 21 and 42 mg of protein in an Avance III 400 MHz ssNMR spectrometer (Bruker) at room temperature. Data was obtained and analyzed using Topspin (Ver. 2.0).

CHAPTER V

BIOPHYSICAL ANALYSIS OF THE TRANSITION OF AN ALL α - HELICAL GREEK-KEY PROTEIN INTO AMYLOID-LIKE FIBRILS COMPOSED OF β -SHEET STRUCTURE

OVERVIEW

In an environment where specific conditions are met, many proteins and peptides can misfold and aggregate into variable length filaments composed of β -sheet structure which can further assemble into intertwined macrostructures [88]. Formation of highly ordered β -sheet aggregates either intracellularly or extracellularly in target tissues, has been associated with many devastating medical disorders such as Alzheimer's and Parkinson's diseases [88, 468]. It is known that the ability to form fibrils is not reliant on the native secondary or tertiary structure composition of the protein and that the sequence of amino acids appears to be fundamental in the fibrillation process and amyloid fibril morphology [181]. Thus there is much research into understanding the role of the amino acid sequence by mutational analysis [179, 208]. The specific mechanisms associated with the formation of β -amyloid fibrils are not yet fully understood. Currently there are two hypotheses describing the mechanism of fibril formation [469]. The first hypothesis includes a nucleation step followed by the elongation step once the nucleus has reached critical mass [469]. The second hypothesis includes the formation of peptide protofibrils of variable length which associate to form fully formed fibrils [469]. The elongation step

is independent of initial fibril formation and has been shown to grow by the binding of monomeric protein to fibril extremities [469]. The process of forming amyloid fibrils *in vitro* typically involves incubating proteins or peptides under one or more extreme conditions such as elevated temperatures, acidic or basic pH, high ionic strength and mechanical agitation [232, 470, 471]. Thus, destabilization of the native state appears to be a pivotal step in the conversion process *in vitro*. Chemical inducers such as metal nanoparticles (TiO₂) and destabilizers (SDS, 2,2,2-trifluoroethanol and heparan sulfate) have also been found to promote fibril formation [237, 472-474].

A key conceptual finding was the realization that proteins unrelated to disease states can be induced to form fibrils [181]. Several proteins not associated with diseases such as the SH3 domain of the p85 α subunit of bovine phosphatidylinositol 3-kinase, acylphosphatase, β -lactoglobulin and myoglobin were induced *in vitro* to form amyloid fibrils [181, 182, 475, 476]. It has thus been hypothesized by Christopher Dobson and co-workers that all proteins have the intrinsic ability to form amyloid fibrils [181]. We are interested in testing this hypothesis with proteins that share the Greek-key topology; one of the most abundant forms within the protein universe [360]. Proteins with this topology can uniquely be composed of different secondary structures and comprise three distinct superfamilies: the all β -sheet immunoglobulins, the mixed α/β -plaits and the all α -helical death domains [360]. Representative proteins from each superfamily have been shown to form fibrils. β 2M from the immunoglobulin superfamily forms naturally occurring fibrils in patients which undergo prolonged renal dialysis [222]. Acylphosphatase and HypF-N are in the α/β -plait superfamily and unlike β 2M, only form

fibrils in an *in vitro* environment [181, 477]. The Apaf-1 CARD protein from the death domain superfamily has also been recently induced to form fibrils *in vitro* [232].

The death domain of the human Fadd-DD from the death domain superfamily contains 104 amino acids arranged into six α -helices which adopt the Greek-key topology (Figure 75A) [405, 478]. It functions in the apoptotic pathway and induces programmed cell death as part of a death-inducing signalling complex [405]. In this chapter, we present the results of comprehensive studies focused on determining the specific conditions required to induce amyloid fibril formation in Fadd-DD and monitored this transition with different spectroscopic techniques. This is a fascinating transition which is most evident in the transformation of a globular, monomeric, helical protein into ordered β -sheet polymers. With over twenty disease states associated with amyloid fibrils there is a clear necessity to fully understand the physiological conditions which lead to the generation of amyloid fibrils and protofibrils so that we may better understand the pathogenesis of amyloidosis [88]. Increased understanding into this mechanism can also inform the design of targeted inhibitors and ultimately lead to viable treatments which do not exist at present. Further investigation of proteins not associated with disease will provide insight into the key determinants of amyloid formation. Understanding how proteins in general evolved to avoid amyloid fibril formation is also pivotal towards resolving the fundamental protein folding problem.

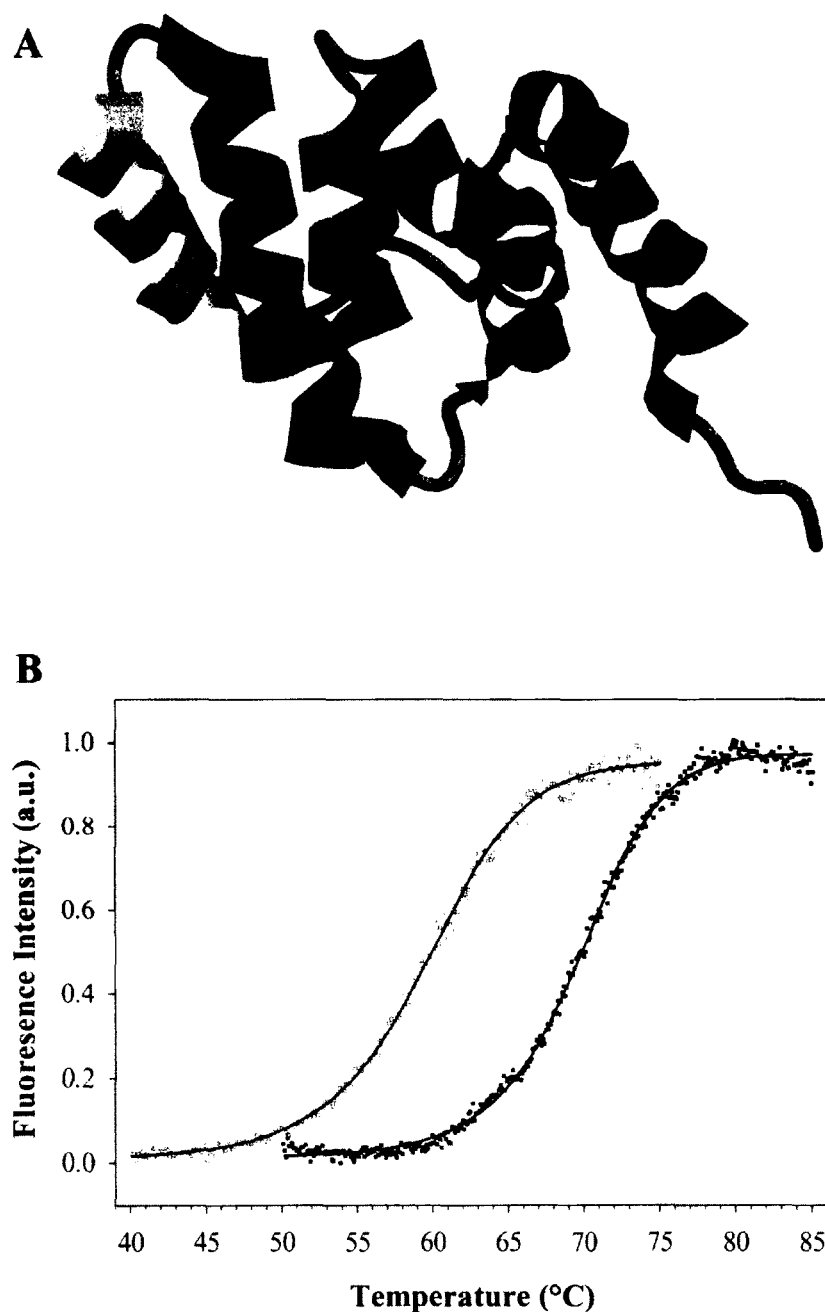


Figure 75. Fadd-DD structure and thermal denaturation assay. (A) Three-dimensional ribbon drawing of the Fadd-DD protein drawn in RasMol (PDB ID: 1E3Y). The helices 1-6 are colored on a scale from blue to red gray. This domain comprises residues 89-192. (B) Thermal unfolding of 0.05 mg/ml Fadd-DD protein in Bis-Tris buffer (pH 6.2) (black squares) and glycine-HCl (pH 2.1) with 150 mM NaCl (gray squares).

RESULTS AND DISCUSSION

In order to elucidate the effect of low pH on the stability of Fadd-DD, thermal unfolding was performed under native conditions (pH 6.2) and in acidic conditions (pH 2.1) (Figure 75B). Analysis of the data revealed that under native conditions Fadd-DD is fully folded and thermo-stable up to a temperature of ~60 °C. The unfolding transition occurs approximately between 60-80 °C with a T_m at 69 °C. At low pH the Fadd-DD protein only remains stable and folded up to ~50 °C and the unfolding transition occurs between approximately 50-68 °C with a T_m of 59 °C. From this data we can see a clear deviation from the native state stability caused by altered solution conditions. This study provides information on the starting conditions to be implemented to induce fibril formation. We initially selected 50 °C at pH 2.1 as the optimal incubation temperature for fibril formation because it lies near the beginning of the transition between folded and unfolded states.

The successful conversion of Fadd-DD to amyloid-like fibrils occurred under the following conditions: 10 mg/ml Fadd-DD in 20 mM glycine-HCl (pH 2.1) and 150 mM NaCl was incubated with shaking at 50 °C and 180 rpm respectively for 278 hours. Ninety-five trials were conducted with a range of conditions which included various buffers such as phosphate-buffered saline, citrate, 2-[N-morpholino]ethanesulfonic acid, and glycine-NaOH, variable protein concentrations and a pH range of 2 to 11, high temperature (60-85 °C), physiological temperature (37 °C), and tests with trifluoroethanol an agent known to induce fibril formation with the amyloid- β peptide and transforming growth factor β induced protein [479]. Several fibril seeding trials were conducted using lysozyme fibrils as a potential inducer because it has been shown that

addition of fibrillar structures can potentially reduce or eliminate the lag phase thereby facilitating amyloid fibril formation [88]. However, these trials were unsuccessful. Increasing and decreasing concentrations of dithiothreitol (DTT) and β -mercaptoethanol (β ME) were used in high pH trials to prevent the three free cysteine residues from forming intermolecular disulfide bonds and precipitating the protein out of the solution. Interestingly, the conversion of an all α -helical protein into only β -amyloid structures occurred in one specific condition out of ninety-five trials. This clearly indicates that the conversion process is not trivial and can only be achieved under very specific conditions.

We monitored the conformational changes from native α -helical structure to the β -amyloid structure using the chromophore ThT. This reagent becomes strongly fluorescent upon specific binding to amyloid fibrils and is considered a standard probe of the lag and growth phases associated with amyloid fibril formation [297]. In Figure 76A we show ThT fluorescence of Fadd-DD samples taken at several time points during the incubation of the protein. An overall increase in ThT intensity indicates the formation of Fadd-DD fibrils. The lag phase for Fadd-DD fibril formation is short and shows signs of a growth phase after only 24 hours of incubation. Growth rapidly continues until approximately 183-254 hours where growth rate of the fibrils is slowed, most likely by reduced availability of the monomeric form of Fadd-DD. A sample was taken at 518 hours and the presence of amyloid fibrils was again verified using CR UV spectroscopy which resulted in an increase in absorbance at 540 nm suggesting amyloid fibril formation (data not shown) [480]. Congo red is a histological dye that has been shown to specifically bind the amyloid form and provide an increase in UV absorbance intensity [481].

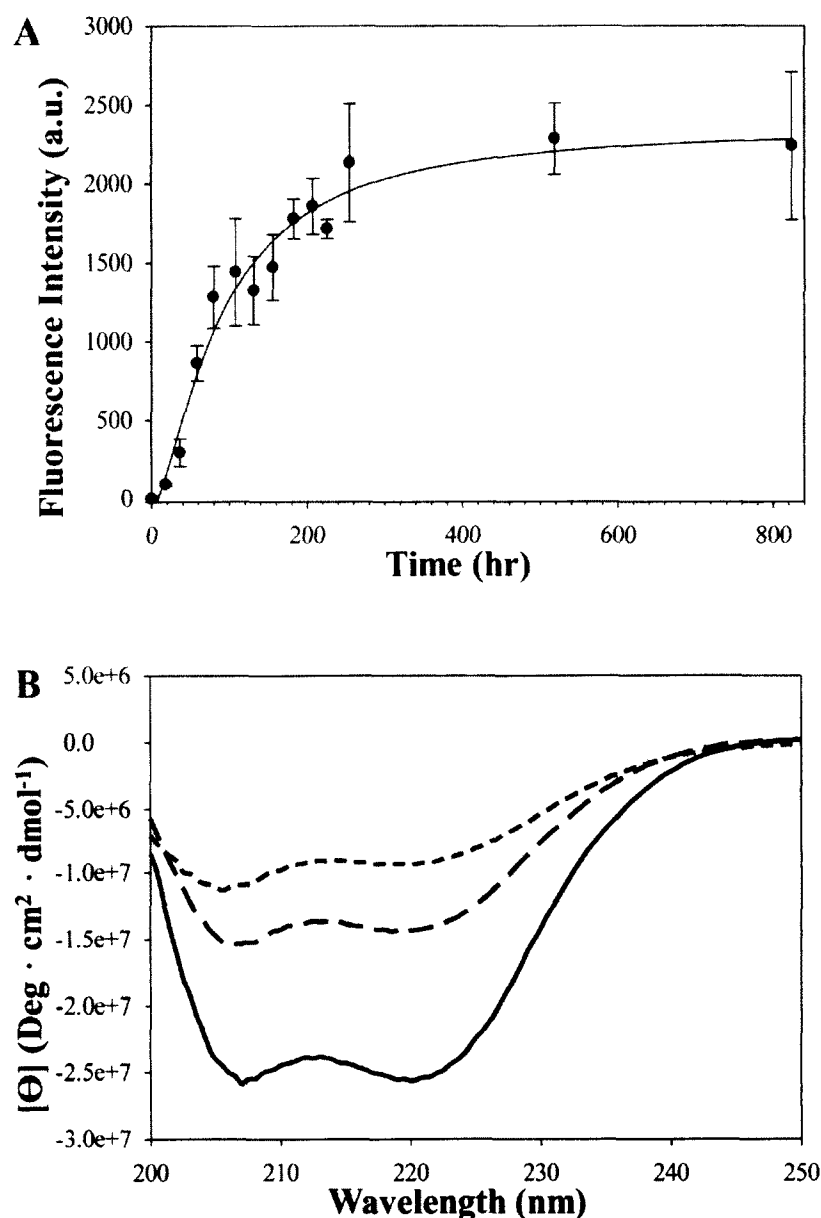


Figure 76. Fibril formation studies using ThT fluorescence and far-UV CD. (A) ThT fluorescence time course of 10 mg/ml Fadd-DD over 823 hours of incubation at pH 2.1, 50 °C and 180 rpm shaking. The average of a triplicate incubation is shown. The data points were fit with a Sigmoidal Logistic, 4 Parameter regression curve in SigmaPlot (ver. 10) (B) Far-UV CD of 0.2 mg/ml of native Fadd-DD in glycine-HCl (pH 2.1) and 150 mM NaCl at 20 °C (solid-line), 35 °C (long-dash line), 50 °C (short-dash line), and 80 °C (dot-dot-dash line).

Using the biophysical technique CD we can investigate structural components of the protein folding process [139]. We monitored the effect of temperature on the secondary structure of native Fadd-DD in 20 mM glycine-HCl, 150 mM NaCl (pH 2.1) using temperature controlled far-UV CD. In Figure 76B we show that the secondary structure of the Fadd-DD protein at 50 °C still maintains a degree of α -helical structure however destabilized, indicated by the molar ellipticity difference. This suggests that at 50 °C (pH 2.1) Fadd-DD has a native-like conformation but is destabilized in accordance with the thermal unfolding data. Using far-UV CD we were able to monitor the secondary structure composition of the natively folded Fadd-DD during the transition from α -helical to β -sheet structure (Figure 77). We monitored the far-UV CD at ~24 h intervals initially and the data indicates a conformational transition of the native α -helical structure to a stable β -sheet composition. Figure 77A shows the CD spectrum of the all α -helical fold of native Fadd-DD. The native fold of Fadd-DD is associated with two minimums around 207 nm and 221 nm shown in the spectrum [405]. Figure 77B-F shows the secondary structuring from α -helical to β -sheet by the loss of the first minimum at 208 nm as well as an overall increase in molar ellipticity. The transition from native structure to β -sheet composition seems to transition through a mixed species of α and β secondary structure conformations as the β -sheet composition increases the α -helical content decreases. Figure 77E shows the secondary structure composition of Fadd-DD incubated for 80 hours in which we clearly observed a predominant conformational change to β -sheet structure associated with the formation of amyloid fibrils. The conversion to β -sheet secondary structure is indicated by a single minimum at around 219 nm. Figure 77F-I maintain the β -sheet structure out to 823 hours as the kinetic process

reaches equilibrium due to the lack of monomeric protein. These results demonstrate that the specific change from α -helical to all β -sheet structure, indicative of amyloid aggregates, has occurred. Interestingly, the near-UV CD spectra for Fadd-DD at pH 6.2 and at pH 2.1 are very similar (data not shown) indicating tertiary structure stability.

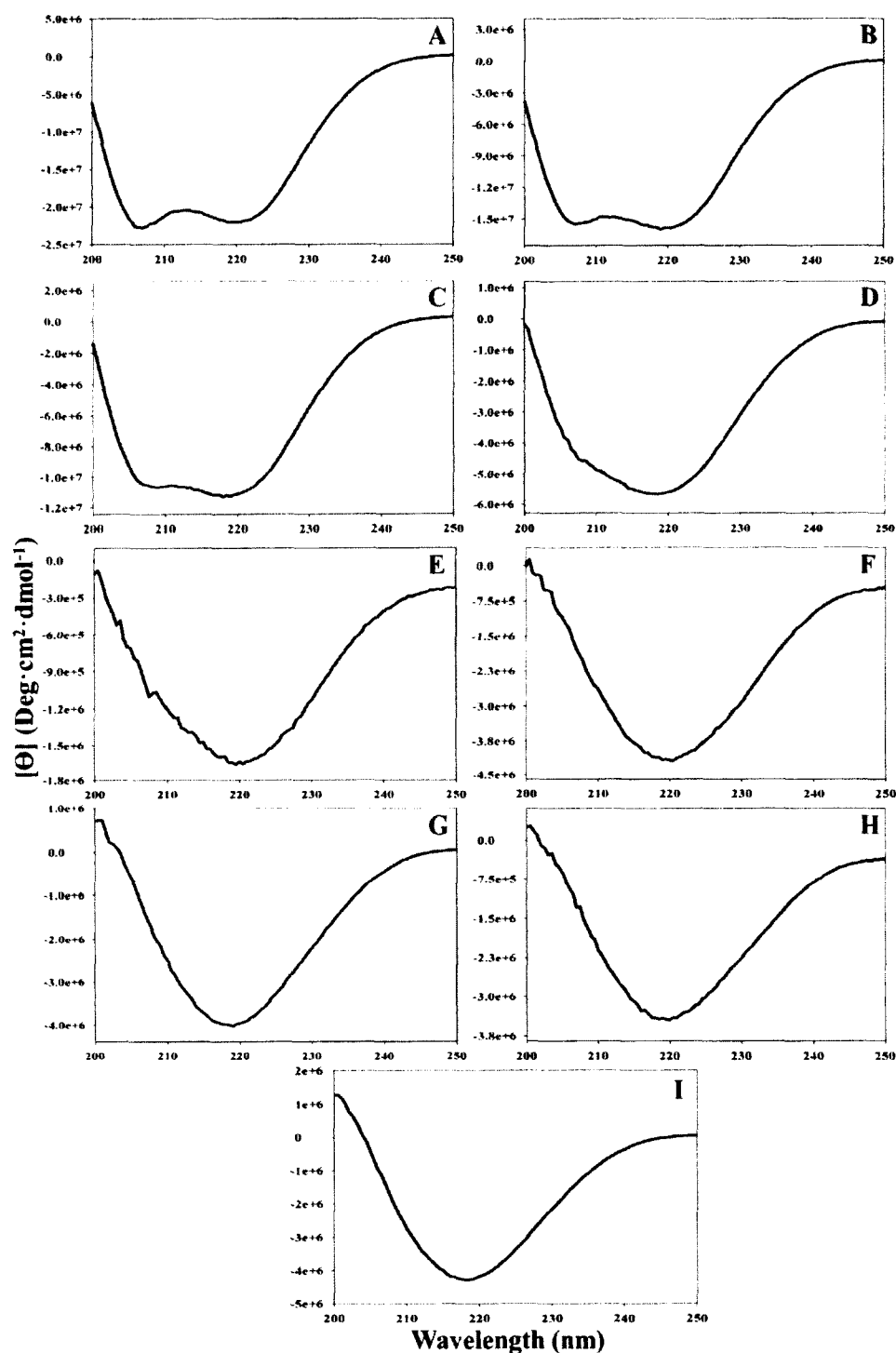


Figure 77. Amyloid fibril formation time course monitored by far-UV CD.

Secondary structure of Fadd-DD measured after (A) 0, (B) 18, (C) 36, (D) 58, (E) 80, (F) 183, (G) 254, (H) 518 and (I) 823 hours of incubation in glycine-HCl (pH 2.1) and 150 mM NaCl buffer at 50 °C and 180 rpm.

Using AFM the formation of Fadd-DD amyloid fibrils was visualized in Figure 78A-I. The formation of Fadd-DD fibrils occurred through the formation of Fadd-DD protofibrils with average sizes ranging from 3-6 nm. After 18 hours the protofibrils were present however at low concentration making it difficult to isolate multiple protofibrils in a single image (Figure 78B). After 36 hours there was an increase in protofibril concentration and elongation appears to have begun, indicated by the solid white arrow in Figure 78C. After 58 hours there is the presence of elongated fibrils with variable lengths as well as protofibrils (Figure 78D). During 80 and up to 823 hours there is an exponential increase of elongated fibrils (Figure 78E-I). Sectional analysis of Fadd-DD fibrils after 80 hours showed two distinct sizes of amyloid fibrils, one ranging from 4-6 nm and the other ranging from 7-9 nm. This indicates the possible presence of a dimeric amyloid fibril containing two amyloid fibril strands wrapped into a single amyloid polymer. The presence of amyloid fibrils was further verified by TEM. After ~11 days, TEM positively established that morphologically short Fadd-DD fibrils were formed (Figure 79). Fadd-DD fibrils tended to have a shorter length in comparison to lysozyme, β -lactoglobulin, myoglobin, and β 2M [182, 222, 237, 476]. Visualization by AFM and TEM under optimized conditions indicates that Fadd-DD fibrils deviate from the “generic” amyloid model, having short fibrils rather than long entangled aggregates.

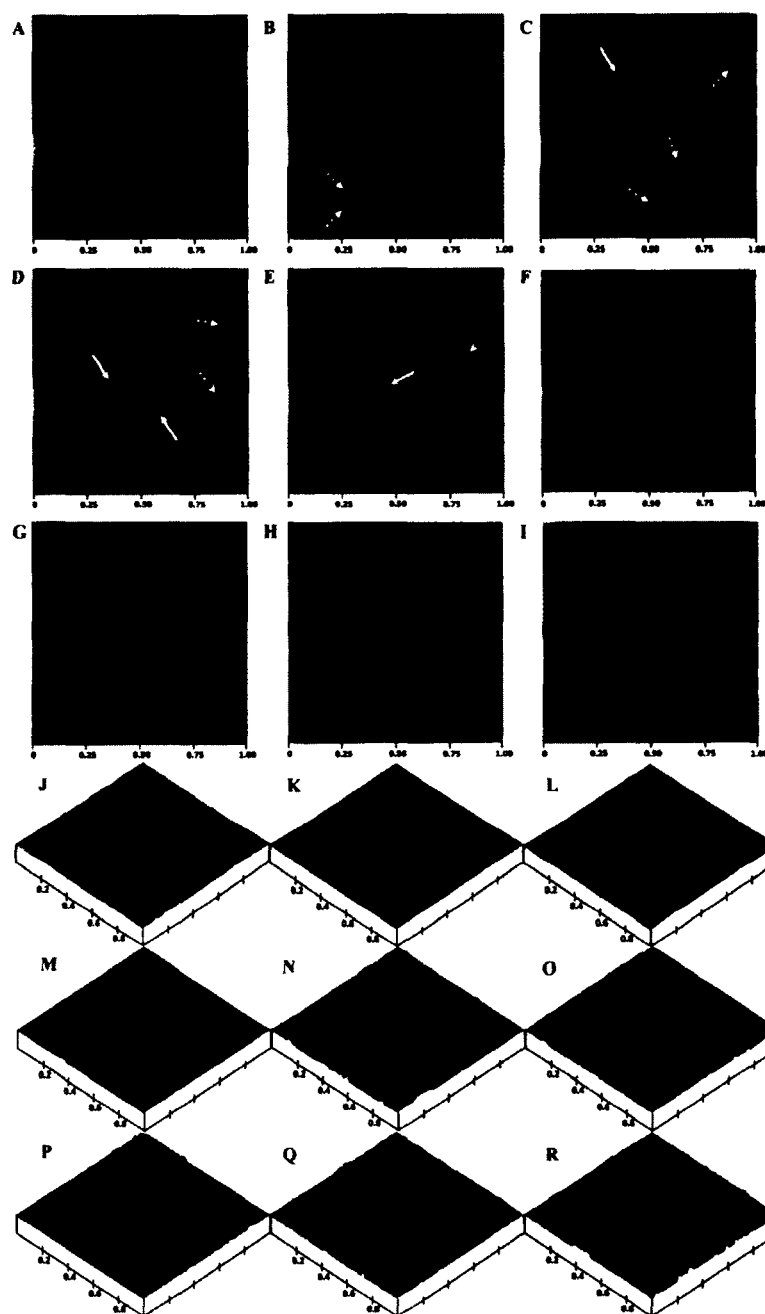


Figure 78. AFM images of Fadd-DD amyloid fibril formation. Topographical images of the formation of amyloid fibrils was done at 0 (**A, J**), 18 (**B, K**), 36 (**C, L**), 58 (**D, M**), 80 (**E, N**), 183 (**F, O**), 254 (**G, P**), 518 (**H, Q**) and 823 (**I, R**) hours. Dotted white arrows indicate protofibrillar species, solid white arrows indicate elongated amyloid species in the 4-6 nm width range and double-line arrow indicate elongated species with in the 7-9 nm width range. Color scales indicate 0-50 nm heights.

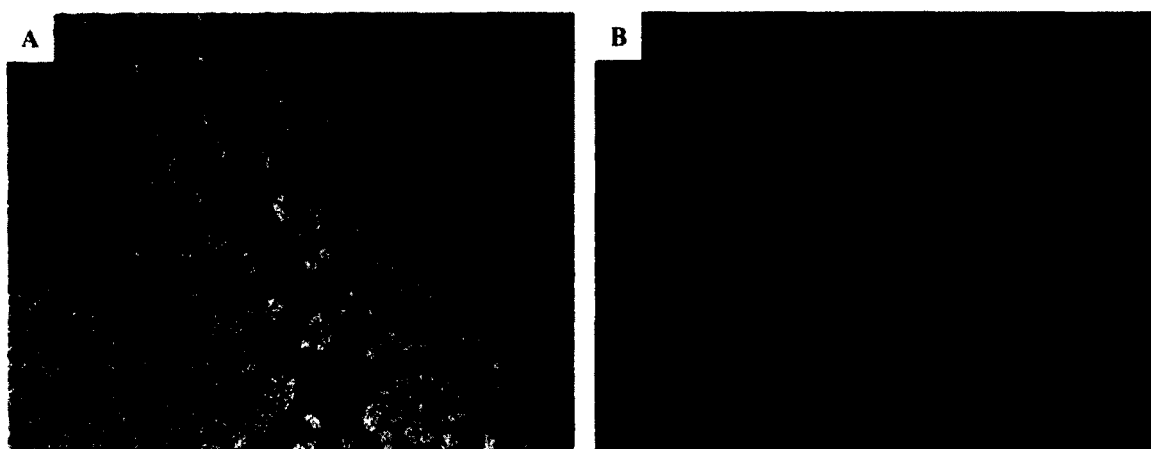


Figure 79. TEM images of Fadd-DD fibrils. (A) Low magnification of Fadd-DD fibril cluster. (B) High magnification of individual Fadd-DD fibril strands. The TEM was conducted with a JEOL JEM-2100F.

It is interesting to note that the morphological deviation observed is related to deviations in environmental factors. We observed that under decreased agitation (75 rpm) both fibrillation and unordered aggregation occur similarly (Figure 80A), whereas under higher agitation (180 rpm) unordered aggregation dramatically decreases and fibrillation becomes more favourable (Figure 80B). Visually the fibrils occurring under decreased agitation appear longer and more similar to fully mature fibrils. Decreasing agitation appears to decrease the formation of protofibrils, increasing the length of the formed fibrils during the elongation. However, conversely under these decreased agitation conditions aggregation also occurs. It appears that there is a competitive reaction occurring between fibril nucleation and unordered aggregation. This suggests that kinetically the pathways of amyloid fibril formation and unordered aggregation under our conditions are close together. Additionally, when decreasing the agitation (rpm) during

fibrillation, nucleation becomes less favourable and unordered aggregation becomes more favourable.

There are several all α -helical proteins reported to form amyloid-like fibrils *in vitro*. These for example include myoglobin, apomyoglobin, the CARD protein and Ure2p prion. Apomyoglobin and myoglobin were induced to form amyloid fibrils similarly in 50 mM sodium borate (pH 9.0), low protein concentrations (1 mg/ml) and were incubated at 65 °C over 24 hours [182, 482]. The Greek-key CARD protein was shown to form fibrils in 50 mM glycine-HCl buffer (pH 2.0) incubated at 60 °C in a heating block [232]. Similarly, Fadd-DD forms fibrils under low pH conditions however deviations in temperature to 50 °C and the addition of NaCl for ionic strength and translational motion were required to overcome the stability of Fadd-DD and induce the fibrillar conformation (Figure 81).

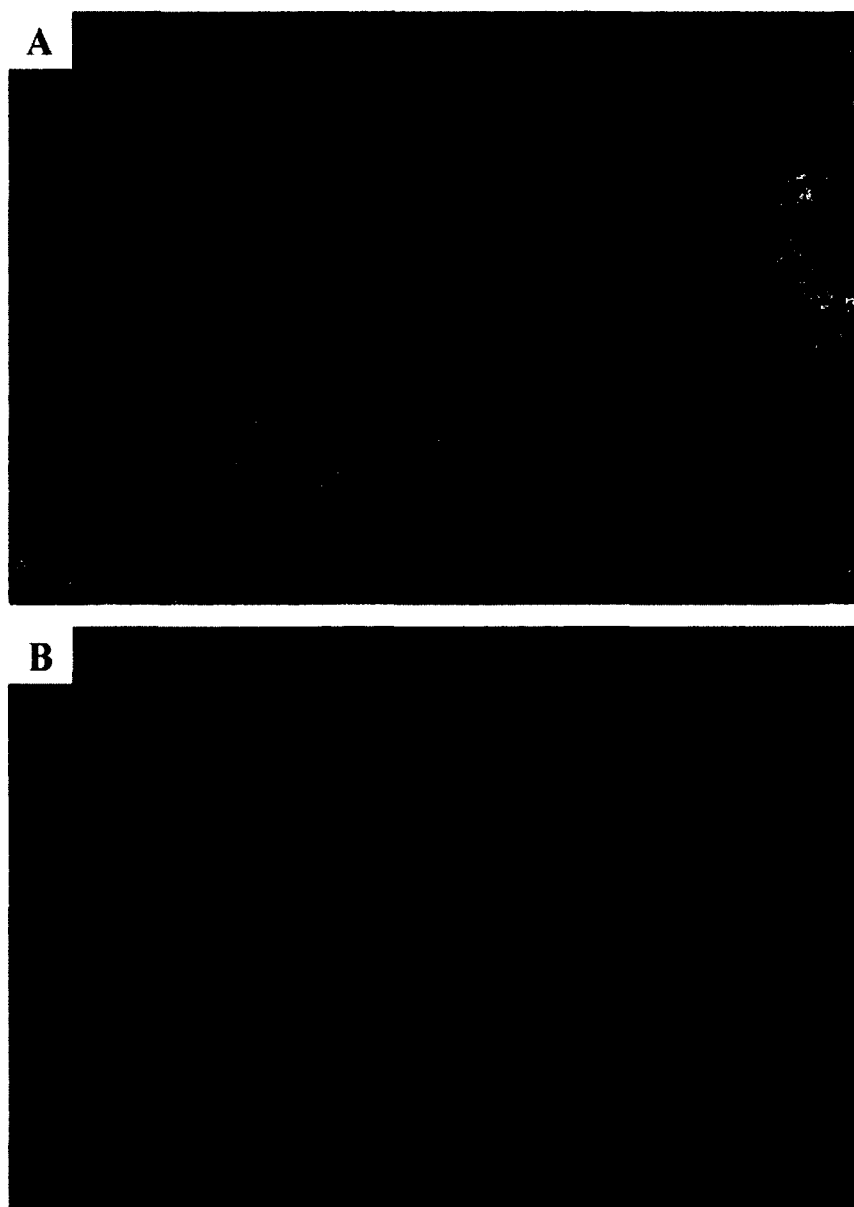


Figure 80. Effects of agitation on Fadd-DD fibril growth. TEM imaging of Fadd-DD fibrils grown in glycine-HCl (pH 2.1) and 150 mM NaCl buffer at 50 °C with agitation of (A) 75 rpm and (B) 180 rpm.

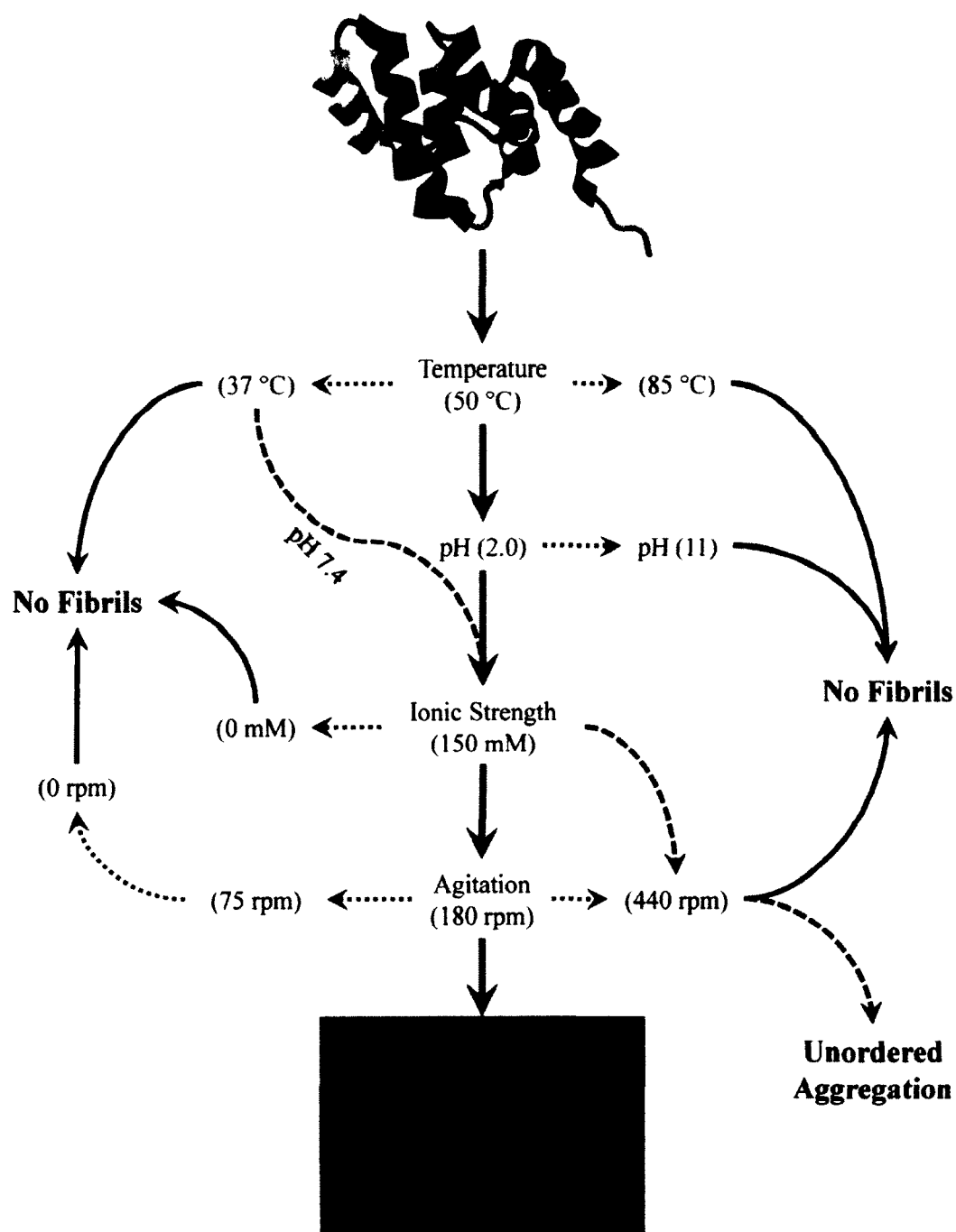


Figure 81. Representative conditions delineating pathways towards and away from amyloid-like fibril formation. Optimized condition for Fadd-DD amyloid formation pathway shown by the green arrows. Dotted red arrows indicate alterations in the conditions pathway and blue arrows indicate resulting effect after incubation. Physiological are conditions shown by the dashed black pathway.

Conversely the all α -helical prion protein, Ure2p, is able to form amyloid fibrils in 50 mM tris (pH 8.4), 200 mM NaCl at 8 °C with no agitation due to a fast assembly property induced by the 18-21 region [483]. Interestingly, Fadd-DD shows no sequence propensity to form amyloid fibrils when analyzed with the Waltz amyloid propensity calculator at pH 7.0 or 2.6 using high specificity (less false positives) parameters [211]. Thus, the environment is the key factor.

The ability to form fibrils under the specific conditions identified in this research supports the “generic amyloid” hypothesis which proposes that formation of amyloid fibril structure is an inherent characteristic of the polypeptide chain, irrespective of the native secondary or tertiary structure composition [88, 181, 234]. While some proteins tend to aggregate more readily than others, there are conditions that can be identified to destabilize the native structure and induce successful conformational exploration into alternative folding states leading to the low energy, β -sheet rich amyloid form [88]. The formation of Fadd-DD fibrils shows a very short lag phase and formation of protofibrillar species. After formation of protofibrils, the formation of amyloid fibrils occurs exponentially during the growth phase. However, it appears that agitation is a critical factor in our model system for altering the kinetics of the nucleation step.

There appears to be no universal set of conditions favourable for amyloid fibril formation. As with protein folding and crystallization, optimal conditions are needed and vary depending on the protein. The results from studies with Fadd-DD presented in this paper clearly indicate that while proteins may have the intrinsic ability to form fibrils they can be very resistant. Previously published papers on this topic do not delineate the failed conditions, unless testing inhibitors or accelerators, which would actually be quite

informative in this relatively new field of study. In our research, only one of ninety-five conditions generated predominantly amyloid fibrils (Figure 81). It is clear however that nature has evolved proteins to avoid fibril formation by optimizing the native state over a fibrillar structure under relevant physiological conditions. The basis for why Fadd-DD does not form fibrils *in vivo* appears to be the need for a specific set of extreme environmental conditions.

MATERIALS AND METHODS

The chemicals used were all high quality reagents. Dithiothreitol was purchased from Fisher (Hanover Park, IL), Bis-Tris from Acros Organics (Geel, Belgium) and ThT was purchased from MP Biomedical (Irvine, CA). Protein was prepared using Ni-NTA His-Bind resin from Novagen (Madison, WI), Sephadex G-75 superfine resin from Sigma (St. Louis, MO) and concentrated with Vivaspin concentrators from Sartorius Stedim Biotech (Bohemia, NY). All buffers used in these studies were filtered through either a 0.45 or 0.22 μ m acrodisc sterile filter (Pall Life Sciences, Suwannee, GA). Microscopy materials (grids and uranyl acetate) were purchased from Electron Microscopy Sciences (Hatfield, PA). The clone was donated by Professor Paul Driscoll (University College London, U.K.). Fadd-DD was expressed and purified in accordance with previously published procedures (IBC biosafety protocol #14-016) [405].

Thermal Unfolding

Fadd-DD protein stock was buffer exchanged into native buffer (20 mM Bis-Tris, 10 mM DTT, pH 6.2) and non-native buffer (20 mM glycine-HCl, 150 mM NaCl, pH 2.1) using a vivaspin 3,000 MWCO PES concentrator at 4 °C and centrifuged at 7,500 rpm in a Rotanta 460R Hettich centrifuge until a concentration of 0.05 mg/ml was achieved for both conditions. Concentration was measured via analysis at 280 nm with a Varian Cary 50 Bio UV spectrophotometer and the calculated extinction coefficient. Using a Varian Cary Eclipse fluorescence spectrophotometer with a peltier thermo controller, fluorescence intensity was measured at 370 and 330 nm over a temperature range of 40-85 °C. The ratios of the two intensities (370 nm /330 nm) was plotted and analyzed using the SigmaPlot (version 10) scientific graphing program (San Jose, CA)

[405]. The midpoint of the unfolding transition (where $\Delta G \approx 0$) was calculated using the following equations [12]:

$$\chi_{\text{unfolded}} = (A_{\text{obs}} - A_{\text{folded}}) / (A_{\text{unfolded}} - A_{\text{folded}})$$

$$K_{\text{eq (folded} \rightarrow \text{unfolded)}} = \chi_{\text{unfolded}} / (1 - \chi_{\text{unfolded}})$$

$$\Delta G = -RT \ln K_{\text{eq}}$$

The χ_{unfolded} is the calculated fraction unfolded based on the observed intensity (A_{obs}).

A_{folded} and A_{unfolded} are the baseline values for fully folded and unfolded states. K_{eq} is the equilibrium ratio between folded and unfolded, based on a normalized scale for the A_{obs} .

Amyloid Fibril Formation

Fadd-DD protein stock was buffer exchanged into 20 mM glycine-HCl 150 mM NaCl, pH 2.1 buffer using a vivaspin 20 ml concentrator and centrifuged at 7,500 rpm until a 10 mg/ml Fadd-DD concentration was reached. 1 ml aliquots of a 10 mg/ml stock of Fadd-DD in a 1.5 ml conical screw cap microcentrifuge tube with O-rings were incubated in a New Brunswick Scientific Innova 4000 shaker incubator at 50 °C and 180 rpm for 278 hours. Fibril formation studies were done in triplicate.

Thioflavin T Fluorescence Assay

A stock solution of 2.5 mM ThT was made with 10 mM potassium phosphate buffer (pH 7.4) and 150 mM NaCl and dissolved overnight [480]. The ThT working solution was prepared by a fifty-fold dilution of the stock solution, filtered through a 0.22 μm sterile filter and stored in the dark. Each time before use a ThT working solution was made fresh as described. Protein samples were collected and diluted forty-fold to a final concentration of 0.25 mg/ml. Samples were analyzed by adding 20 μl of the dilute protein sample to 200 μl /well of working solution in a black 96-well microtiter

fluorescence plate. Microtiter plates were analyzed in a BioTek Synergy HT microtiter plate reader from BioTek Instruments Inc. (Winooski, VT). The parameters for the experiment are as follows: $\lambda_{\text{ex}} = 440 \text{ nm}$ and $\lambda_{\text{em}} = 485 \text{ nm}$. The average of a triplicate study is graphed using SigmaPlot (version 10).

Circular Dichroism Analysis

Fadd-DD was diluted to a final concentration of 0.25 mg/ml for far-UV CD from a 10 mg/ml stock solution. Fadd-DD fibrils samples were diluted similarly for far-UV CD and the protein concentration was measured by UV-Vis absorption spectroscopy at 280 nm with the calculated extinction coefficient. The WT-Fadd-DD and fibril samples were analyzed on a MOS 450 (Bio-logic, France). Far-UV CD was conducted with 1 mm path length quartz CD cuvette and slit widths of 1 mm. The temperature was maintained at 20 °C. 15 to 30 spectra are averaged and graphed using a SigmaPlot (version 10).

Atomic Force Microscopy Imaging

Samples from far-UV CD were immediately rerun on an atomic force microscope. 5 μl of the dilute fibrils were pipetted onto a cleaned glass sample holder and spread to an approximate 0.25 cm diameter and allowed to stand for 10 minute before excess solution was wicked from the surface. The sample was allowed to air dry and then was washed 4-5 times with 20 μl of distilled water and allowed to stand for 1 minute before similarly removing the excess solution and air drying. Samples were imaged on a Veeco DiNanoscope 3D AFM using AM mode microscopy.

Transmission Electron Microscopy

Fadd-DD fibril solutions were diluted 5-fold in distilled water. 10 μl of Fadd-DD fibrils were pipetted onto a 400 mesh formvar carbon grid and allowed to stand for 1

minute. Excess sample was wicked off and the grid was allowed to air dry before 20 μ l of a 2% uranyl acetate stock solution was applied to negatively stain the Fadd-DD fibrils for 30 s. After 30 s excess urinal acetate was wicked off and samples were immediately rinsed with distilled water. After rinsing, the sample was heated to 75 °C overnight before storage in desiccators until imaging at the Applied Research Center (Newport News, VA) on a JEOL JEM-2100F field emission TEM at 200 kV.

CHAPTER VI

CONCLUSIONS AND FUTURE WORK

This dissertation is an investigation into understanding and further enhancing our knowledge of protein folding and misfolding. Protein folding is fundamentally important if we hope to develop therapeutics and combat protein misfolding diseases. It is also an integral unanswered problem in science. Its resolution is key to advancing our knowledge of one of the most basic and foundational biological processes. Proteins were studied using a multitude of methods and techniques. In this work we investigated protein folding using computational and experimental methods aimed at developing our understanding of protein evolution and how the amino acid sequence dictates the final topology, respectively. In the protein folding section (Chapters II, III and IV) of this work we have focused on GB1 a small 56 amino acid IgG-binding protein. We investigated the evolutionary relationship between GB1 and the GA module, determined structurally conserved residues in GB1 as well as the effect of a point mutation on the structure and folding of GB1. We also used GB1 as our model system in developing a novel method to follow the formation of a conserved set of long-range interactions between Phe30, Phe52 and Trp43 in real-time using NMR. In the protein misfolding section (Chapter V) we explored the concept of fibrillation being an alternative folding pathway and the idea that every protein has the ability to form the fibril morphology under a certain set of conditions. We used Fadd-DD as our model system and determine the specific conditions required in converting this all α -helical Greek-key protein into amyloid-like fibrils.

Evolutionary Analysis of GB1 and GA

GB1 and GA are small domains with differing topology and functions. GB1 is a $4\beta+\alpha$ fold and the IgG-binding domain of protein G, whereas GA is a 3α -helical bundle fold and the albumin-binding domain of protein L. He *et al.* showed that these two proteins could be mutated into a sequence of up to 98 % identity and still maintain their unique folds. With a point mutation these two proteins were able to switch folds and functions. Being able to switch folds is a unique process in evolution and could be related to the evolution of these two proteins. Initial evidence suggests that due to their low sequence identity GB1 and GA did not evolve from each other directly, but rather from another group of proteins. Further research indicated that these two proteins were uniquely related and have evolved from a common ancestral family of proteins. In our work we used bioinformatics techniques to explore the evolution of the WT-GB1 and WT-GA. Using PSI-BLAST, an iterative process for determining distant relationships we discovered a relationship between WT-GB1 and WT-GA to a group of YSIRK signalling proteins, respectively. These two proteins, YSIRK signalling, were not interrelated but further analysis showed that they were both related through a group of mucus-binding proteins.

These proteins may have evolved from a duplication event, followed by subsequent mutations, sequence recombination or translocation giving rise to a new structure and function. The sequence alignment of our related proteins with the mucus-binding protein showed a unique overlap of the common sequence, LINxxKTV, similar to what could be considered a motif. Interestingly, instead of evolving this sequence independently of each other it seems that nature has simply reused it to evolve another

protein sequence in a more efficient manner. The low variation associated with this short evolutionary sequence motif indicates that it is probably critical to both proteins and was conserved through evolution. Phylogenetic, sequence motif and MD data all seemed to suggest that GB1 and GA evolved from the mucus-binding protein through a YSIRK related evolutionary intermediate. Interestingly, a characterized mucus-binding protein was shown to have residual Ig-binding function which strongly supports our data.

In the future we intend to further explore the structures of the YSIRK and mucus-binding proteins using other modelling methods such as *ab initio* modelling in order to more confidently understand the structure and possible functions of these proteins. In addition it would be interesting to look at how the native 61 amino acid GA module fits into this evolutionary tree. We only intended to expand upon and use the two evolved sequences shown to swap structures with high sequence identity (GA and GB1), in this present dissertation research.

Conservation, Structural and Kinetic Analysis of GB1

To study the structure and folding of GB1 we needed to get a clearer understanding of which residues were more important for folding. To determine those amino acids we performed a detailed bioinformatics conservation analysis using structural comparison. We determined that there were 9 correctly conserved residues in the sequence of GB1 with 8 being predominately hydrophobic and 1 hydrophilic in nature. Interestingly, there was at least one conserved residue found in each secondary structure component which may be important in maintaining the $4\beta+\alpha$ fold and also may be critical to the thermal stability of GB1. To gain insight into the effect of a subtle

mutation such as Phe52Tyr, we used site-directed mutagenesis to change the sequence and characterized the resulting structure using biophysical techniques. It seems that this mutation conferred increased secondary structure content but reduced the structural stability of GB1. However, this change did not seem to affect the ability of GB1 to fold to its native conformation. This result was important as we were looking to remove the second Phe from the structure. For clarity, this mutation and characterization was conducted as preliminary experimentation focused on its applicability to future work done in Chapter IV. However, due to the reduced stability and structural effects observed we opted to proceed with the WT-GB1 form for our next research studies.

Development of a Method to Monitor Long-range Interactions in GB1 using NMR

The goal of this project was to develop preliminary methodology for a novel folding-freezing method in which a kinetic intermediate is captured by flash freezing during the folding process using supercooled isopentane. GB1 was chosen as our model system and we selected long-range interactions between Phe30, Phe 52 and Trp43 in the core of GB1. To assess the applicability of GB1 to our folding-freezing method coupled with MAS-NMR we needed to experimentally approach the problem from two avenues. First we assessed the effect on WT-GB1 folding at high concentrations which would be required for capturing a kinetic intermediate at concentrations visible to a NMR spectrometer. Thus, we characterized the folding of WT-GB1 and Phe52Tyr-GB1 at high concentrations (~100 mg/ml), which revealed that there were no adverse effects observed during the folding at high concentrations. We then synthesized the selectively labelled protein using our Aux(DE3) system. Synthesizing a ^{13}C -Phe/ ^{13}C -Trp uniquely labelled

WT-GB1 protein using Aux(DE3) bacterium was a significant achievement as the methodology needed to be refined and controlled to ensure proper growth and selective labelling. We were able to successfully synthesize over 500 mg of ^{13}C -Phe/ ^{13}C -Trp labelled WT-GB1. This was a major accomplishment as we were able to maximize the use of our isotopically labelled amino acids. Additionally, we synthesized a significant amount of uniformly ^{13}C -labeled WT-GB1 for control studies.

The second avenue involved experimentation with NMR. There we tested our ^{13}C - ^{13}C DARR pulse program using uniformly ^{13}C -labeled WT-GB1 and ^{13}C -Phe/ ^{13}C -Trp labelled WT-GB1. We performed a DARR experiment on lyophilized ^{13}C -Phe/ ^{13}C -Trp labelled WT-GB1 and were able to see CRPK's of interest. Interestingly, we were only able to see two of the four CRPK's we were anticipating from the distances calculated from the structure. This could indicate that the lyophilized form of our protein may have differences in structure compared to what was found in the hydrated state. We further explored this using MAS-NMR with the same protein under solution-state conditions followed by lyophilization directly in the rotor and reassessing the structure. Uniformly ^{13}C -labeled WT-GB1 resulted in a clustering of many peaks and proved to be less amenable to the DARR experiment. However, we needed ~41 mg of protein to quickly see our CRPK's of interest and in the future we need to further assess the concentration dependence of ^{13}C -Phe/ ^{13}C -Trp labelled WT-GB1. We need to know what the minimal concentration of ^{13}C -Phe/ ^{13}C -Trp labelled WT-GB1 is required to see well-resolved CRPK's and the minimal required acquisition time. An important control would involve analysing unlabelled protein for natural abundance signal in comparison to our enriched protein. This would be critical in evaluating whether chemical denaturation is

the most applicable method for the folding-freezing method. Additionally, these foundational studies also suggest thermal unfolding and refolding may be another avenue to explore, since higher protein quantities can be used without dilutions.

Due to the concentration related issues we began characterizing an alternative folding method by solution-state NMR using pH changes to denature and fold our protein as this would require a significant reduction in dilution conditions for refolding. In addition, this was an opportunity to explore a possible molten globule structure, which are typically inducible at low pH's. We analysed ^{13}C -Phe/ ^{13}C -Trp labelled WT-GB1 at pH 2, 7 and 12 using 2D ^1H , ^{13}C correlation HSQC. We found that even under pH 2 and 12 the environment of the labelled Phe and Trp remained relatively unchanged. Interestingly, CD indicated there was a significant loss in secondary structure content which was quite surprising. Our CD and HSQC data indicated that we may have found an intermediate-like state containing native-like tertiary structure with very little secondary content which could be a reverse molten globule. We propose that it may be a conserved expanded native topology (CENT). However, we will need to explore this potentially ground breaking result further in the future. It is reasonably straightforward to explore the hydrogen-bonding of this new state by NMR using ^{13}C - ^{15}N labelled WT-GB1. MD can also be applied to assist in building a model of the proposed CENT structure.

Formation of Amyloid-like Fibrils with Fadd-DD

The propensity for proteins or peptides to misfold into amyloid deposits is associated with many neurodegenerative and systemic amyloid related diseases which include most notably Alzheimer's and Parkinson's. It was originally thought that

information stored in a polypeptide chain only encodes for a single native fold. The transition of a protein from its soluble functional form into the β -sheet rich insoluble ordered filamentous polymer is a special form of misfolding and possess a great intellectual challenge to the scientific community which seeks to understand the underlying mechanism. It is now believed that amyloid fibrillation may be a separate alternative folding pathway. This process is not unique to any specific class of proteins and appears to be a unique alternative conformation of the polypeptide backbone, in which fibrillation conditions are dictated by the amino acid sequence.

The tendency for a protein to alternatively form highly ordered amyloid fibrils is dependent on many biological factors. Mutations, temperature, concentration, translational motion and pH play a pivotal role in inducing fibril aggregate assembly *in vitro*. The key feature appeared to be the need to destabilize the native state structure as a required first step. In this chapter we reported on the detailed conversion of the death domain of the human Fadd-DD, an all α -helical protein with a Greek-key topology, into an all β -sheet amyloid-like fibril, using a comprehensive range of spectroscopic techniques that provide insight into this process. This transition from α -helical to β -sheet seems to require destabilization but not complete loss of the secondary structure to explore alternative conformations. This was a fascinating transition that supports the hypothesis that all proteins have the innate ability to form a fibril-like structure. Thus, the primary structure can encode two alternative 3D structures: the native, functional state and the β -amyloid state. We investigated the specific environment required to convert the Fas-associated death domain into an amyloid-like conformation. Fas-associated death domain does not appear to naturally form amyloid fibrils *in vivo*. Our results clearly

indicate that proteins evolved to avoid amyloid fibril formation because we find that the conditions required for formation in our model system are very specific and far from physiological. There is still much that can be done using the Fadd-DD fibrillation system we developed. We can explore mutations in the Fadd-DD sequence to see if there are mutations that will allow the formation of Fadd-DD fibrils at physiological conditions. This would be important in understanding how amyloid related fibrils can form under physiological conditions. It is also important to conduct ssNMR studies of isotopically labelled protein to solve the 3D structure of Fadd-DD fibrils.

In summary, protein folding and misfolding are two sides of the same coin that is foundational to advancing not only scientific knowledge but medical research, therapeutics and disease prevention. Discovering new forms and folds can lead to revolutionary alternative avenues of research. As instrumentation and scientific methods rapidly advance, the importance of solving more complex protein structures with high resolution, folding dynamics and functions in of both folded and misfolding protein forms is critical to furthering our understanding of biological life. This area of science is hypothesis driven, experimentally proven and in some cases seems to require a bit of serendipity and in the words of Louis Pasteur, “Fortune favours the prepared mind.”

REFERENCES

1. Fersht A: **Structure and mechanism in protein science : a guide to enzyme catalysis and protein folding**. New York: W.H. Freeman; 1999.

2. Gregory SG, Barlow KF, Mclay KE, Kaul R, Swarbreck D, Dunham A, Scott CE, Howe KL, Woodfine K, Spencer CCA *et al*: **The DNA sequence and biological annotation of human chromosome 1**. *Nature* 2006, **441**(7091):315-321.

3. Collins FS, Lander ES, Rogers J, Waterston RH, Conso IHGS: **Finishing the euchromatic sequence of the human genome**. *Nature* 2004, **431**(7011):931-945.

4. Schmutz J, Wheeler J, Grimwood J, Dickson M, Yang DJ, Caoile C, Bajorek E, Black S, Chan YM, Denys M *et al*: **Quality assessment of the human genome sequence**. *Nature* 2004, **429**(6990):365-368.

5. Dunham I, Shimizu N, Roe BA, Chisoe S, Dunham I, Hunt AR, Collins JE, Bruskiewich R, Beare DM, Clamp M *et al*: **The DNA sequence of human chromosome 22**. *Nature* 1999, **402**(6761):489-495.

6. Levintha.C: **Are There Pathways for Protein Folding**. *J Chim Phys Pcb* 1968, **65**(1):44-&.

7. Zwanzig R, Szabo A, Bagchi B: **Levinthal's paradox**. *Proc Natl Acad Sci U S A* 1992, **89**(1):20-22.

8. Daggett V, Fersht AR: **Is there a unifying mechanism for protein folding?** *Trends Biochem Sci* 2003, **28**(1):18-25.

9. Anfinsen CB: **Principles that govern the folding of protein chains**. *Science* 1973, **181**(4096):223-230.

10. Anfinsen CB, Scheraga HA: **Experimental and theoretical aspects of protein folding**. *Adv Protein Chem* 1975, **29**:205-300.

11. Nolting B, Andert K: **Mechanism of protein folding**. *Proteins* 2000, **41**(3):288-298.

12. Nolting B: **Protein folding kinetics : biophysical methods**, 2nd edn. Berlin ; New York: Springer; 2006.
13. Karplus M, Weaver DL: **Protein folding dynamics: the diffusion-collision model and experimental data**. *Protein Sci* 1994, **3**(4):650-668.
14. Udgaonkar JB, Baldwin RL: **NMR evidence for an early framework intermediate on the folding pathway of ribonuclease A**. *Nature* 1988, **335**(6192):694-699.
15. Kragelund BB, Osmark P, Neergaard TB, Schiodt J, Kristiansen K, Knudsen J, Poulsen FM: **The formation of a native-like structure containing eight conserved hydrophobic residues is rate limiting in two-state protein folding of ACBP**. *Nat Struct Biol* 1999, **6**(6):594-601.
16. Myers JK, Oas TG: **Preorganized secondary structure as an important determinant of fast protein folding**. *Nat Struct Biol* 2001, **8**(6):552-558.
17. Lu JR, Dahlquist FW: **Detection and Characterization of an Early Folding Intermediate of T4 Lysozyme Using Pulsed Hydrogen-Exchange and 2-Dimensional Nmr**. *Biochemistry* 1992, **31**(20):4749-4756.
18. Mullins LS, Pace CN, Raushel FM: **Investigation of Ribonuclease-T(1) Folding Intermediates by Hydrogen-Deuterium Amide Exchange 2-Dimensional Nmr-Spectroscopy**. *Biochemistry* 1993, **32**(24):6152-6156.
19. Brown JE, Klee WA: **Helix-Coil Transition of Isolated Amino Terminus of Ribonuclease**. *Biochemistry* 1971, **10**(3):470-&.
20. Bierzynski A, Kim PS, Baldwin RL: **A Salt Bridge Stabilizes the Helix Formed by Isolated C-Peptide of Rnase-A**. *Proc Natl Acad Sci* 1982, **79**(8):2470-2474.
21. Shoemaker KR, Kim PS, Brems DN, Marqusee S, York EJ, Chaiken IM, Stewart JM, Baldwin RL: **Nature of the Charged-Group Effect on the Stability of the C-Peptide Helix**. *Proc Natl Acad Sci* 1985, **82**(8):2349-2353.
22. Epand RM, Scheraga HA: **The influence of long-range interactions on the structure of myoglobin**. *Biochemistry* 1968, **7**(8):2864-2872.

23. Gay GD, Ruizsanz J, Neira JL, Itzhaki LS, Fersht AR: **Folding of a Nascent Polypeptide-Chain in-Vitro - Cooperative Formation of Structure in a Protein Module.** *Proc Natl Acad Sci* 1995, **92**(9):3683-3686.
24. Rackovsky S, Scheraga HA: **Hydrophobicity, Hydrophilicity, and Radial and Orientational Distributions of Residues in Native Proteins.** *Proc Natl Acad Sci* 1977, **74**(12):5248-5251.
25. Dill KA: **Theory for the Folding and Stability of Globular-Proteins.** *Biochemistry* 1985, **24**(6):1501-1509.
26. Dill KA: **Dominant Forces in Protein Folding.** *Biochemistry* 1990, **29**(31):7133-7155.
27. Akiyama S, Takahashi S, Ishimori K, Morishima I: **Stepwise formation of alpha-helices during cytochrome c folding.** *Nat Struct Biol* 2000, **7**(6):514-520.
28. Agashe VR, Shastry MCR, Udgaonkar JB: **Initial Hydrophobic Collapse in the Folding of Barstar.** *Nature* 1995, **377**(6551):754-757.
29. Nolting B, Golbik R, Neira JL, SolerGonzalez AS, Schreiber G, Fersht AR: **The folding pathway of a protein at high resolution from microseconds to seconds.** *Proc Natl Acad Sci* 1997, **94**(3):826-830.
30. Fersht AR: **Optimization of Rates of Protein-Folding - the Nucleation-Condensation Mechanism and Its Implications.** *Proc Natl Acad Sci* 1995, **92**(24):10869-10873.
31. Itzhaki LS, Otzen DE, Fersht AR: **The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding.** *J Mol Biol* 1995, **254**(2):260-288.
32. Fersht AR: **Nucleation mechanisms in protein folding.** *Curr Opin Struct Biol* 1997, **7**(1):3-9.
33. Uversky VN, Fink AL: **The chicken-egg scenario of protein folding revisited.** *FEBS Lett* 2002, **515**(1-3):79-83.

34. Jackson SE, Fersht AR: **Folding of Chymotrypsin Inhibitor-2 .1. Evidence for a 2-State Transition.** *Biochemistry* 1991, **30**(43):10428-10435.
35. Otzen DE, Itzhaki LS, Elmasry NF, Jackson SE, Fersht AR: **Structure of the Transition-State for the Folding/Unfolding of the Barley Chymotrypsin Inhibitor-2 and Its Implications for Mechanisms of Protein-Folding.** *Proc Natl Acad Sci* 1994, **91**(22):10422-10425.
36. Grant T, Greene L: **A Network Approach To Model Protein Folding.** *Protein Sci* 2012, **21**:226-227.
37. Clarke J, Hamill SJ, Johnson CM: **Folding and stability of a fibronectin type III domain of human tenascin.** *J Mol Biol* 1997, **270**(5):771-778.
38. Matagne A, Radford SE, Dobson CM: **Fast and slow tracks in lysozyme folding: Insight into the role of domains in the folding process.** *J Mol Biol* 1997, **267**(5):1068-1074.
39. Fulton KF, Main ERG, Daggett V, Jackson SE: **Mapping the interactions present in the transition state for unfolding/folding of FKBP12.** *J Mol Biol* 1999, **291**(2):445-461.
40. White GWN, Gianni S, Grossmann JG, Jemth P, Fersht AR, Daggett V: **Simulation and experiment conspire to reveal cryptic intermediates and a slide from the nucleation-condensation to framework mechanism of folding.** *J Mol Biol* 2005, **350**(4):757-775.
41. Travaglini-Allocatelli C, Ivarsson Y, Jemth P, Gianni S: **Folding and stability of globular proteins and implications for function.** *Curr Opin Struct Biol* 2009, **19**(1):3-7.
42. Jackson SE, Fersht AR: **Folding of Chymotrypsin Inhibitor-2 .2. Influence of Proline Isomerization on the Folding Kinetics and Thermodynamic Characterization of the Transition-State of Folding.** *Biochemistry* 1991, **30**(43):10436-10443.
43. Nolting B: **Analysis of the folding pathway of chymotrypsin inhibitor by correlation of phi-values with inter-residue contacts.** *J Theor Biol* 1999, **197**(1):113-121.

44. Shoemaker BA, Wolynes PG: **Exploring structures in protein folding funnels with free energy functionals: the denatured ensemble.** *J Mol Biol* 1999, **287**(3):657-674.
45. Shoemaker BA, Wang J, Wolynes PG: **Exploring structures in protein folding funnels with free energy functionals: the transition state ensemble.** *J Mol Biol* 1999, **287**(3):675-694.
46. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG: **Funnels, pathways, and the energy landscape of protein folding: a synthesis.** *Proteins* 1995, **21**:167-195.
47. Onuchic JN, Luthey-Schulten Z, Wolynes PG: **Theory of protein folding: the energy landscape perspective.** *Annu Rev Phys Chem* 1997, **48**:545-600.
48. Onuchic JN, Wolynes PG: **Theory of protein folding.** *Curr Opin Struct Biol* 2004, **14**(1):70-75.
49. Oliveberg M, Wolynes PG: **The experimental survey of protein-folding energy landscapes.** *Q Rev Biophys* 2005, **38**(3):245-288.
50. Kelly SM, Price NC: **The application of circular dichroism to studies of protein folding and unfolding.** *Biochim Biophys Acta* 1997, **1338**(2):161-185.
51. Mittag T, Forman-Kay JD: **Atomic-level characterization of disordered protein ensembles.** *Curr Opin Struct Biol* 2007, **17**(1):3-14.
52. Ptitsyn O: **How molten is the molten globule?** *Nat Struct Biol* 1996, **3**(6):488-490.
53. Ptitsyn OB: **Molten globule and protein folding.** *Adv Protein Chem* 1995, **47**:83-229.
54. Arai M, Kuwajima K: **Role of the molten globule state in protein folding.** *Adv Protein Chem* 2000, **53**:209-282.

55. Kobashigawa Y, Demura M, Koshiha T, Kumaki Y, Kuwajima K, Nitta K: **Hydrogen exchange study of canine milk lysozyme: stabilization mechanism of the molten globule.** *Proteins* 2000, **40**(4):579-589.
56. Koshiha T, Yao M, Kobashigawa Y, Demura M, Nakagawa A, Tanaka I, Kuwajima K, Nitta K: **Structure and thermodynamics of the extraordinarily stable molten globule state of canine milk lysozyme.** *Biochemistry* 2000, **39**(12):3248-3257.
57. Greene LH, Wijesinha-Bettoni R, Redfield C: **Characterization of the molten globule of human serum retinol-binding protein using NMR spectroscopy.** *Biochemistry* 2006, **45**(31):9475-9484.
58. Redfield C: **NMR studies of partially folded molten-globule states.** *Methods Mol Biol* 2004, **278**:233-254.
59. Redfield C: **Using nuclear magnetic resonance spectroscopy to study molten globule states of proteins.** *Methods* 2004, **34**(1):121-132.
60. Di Paolo A, Balbeur D, De Pauw E, Redfield C, Matagne A: **Rapid collapse into a molten globule is followed by simple two-state kinetics in the folding of lysozyme from bacteriophage lambda.** *Biochemistry* 2010, **49**(39):8646-8657.
61. Rosner HI, Redfield C: **The human alpha-lactalbumin molten globule: comparison of structural preferences at pH 2 and pH 7.** *J Mol Biol* 2009, **394**(2):351-362.
62. Hartl FU, Hayer-Hartl M: **Converging concepts of protein folding in vitro and in vivo.** *Nat Struct Mol Biol* 2009, **16**(6):574-581.
63. Garrett R, Grisham CM: **Biochemistry**, 5th edn. Belmont, CA: Brooks/Cole, Cengage Learning; 2013.
64. Wolynes PG: **Folding funnels and energy landscapes of larger proteins within the capillarity approximation.** *Proc Natl Acad Sci U S A* 1997, **94**(12):6170-6175.

65. Finkelstein AV, Badretdinov A: **Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold.** *Fold Des* 1997, **2**(2):115-121.
66. Finkelstein AV: **Rate of Beta-Structure Formation in Polypeptides.** *Proteins-Structure Function and Genetics* 1991, **9**(1):23-27.
67. Wolynes PG: **Symmetry and the energy landscapes of biomolecules.** *Proc Natl Acad Sci U S A* 1996, **93**(25):14249-14255.
68. Klimov DK, Thirumalai D: **Factors governing the foldability of proteins.** *Proteins* 1996, **26**(4):411-441.
69. Thirumalai D: **From Minimal Models to Real Proteins - Time Scales for Protein-Folding Kinetics.** *J Phys I* 1995, **5**(11):1457-1467.
70. Abkevich VI, Gutin AM, Shakhnovich EI: **Impact of Local and Nonlocal Interactions on Thermodynamics and Kinetics of Protein-Folding.** *J Mol Biol* 1995, **252**(4):460-471.
71. Gutin AM, Abkevich VV, Shakhnovich EI: **Chain Length Scaling of Protein Folding Time.** *Phys Rev Lett* 1996, **77**(27):5433-5436.
72. Sali A, Shakhnovich E, Karplus M: **How Does a Protein Fold.** *Nature* 1994, **369**(6477):248-251.
73. Onuchic JN, Wolynes PG, Lutheyschulten Z, Socci ND: **Toward an Outline of the Topography of a Realistic Protein-Folding Funnel.** *Proc Natl Acad Sci* 1995, **92**(8):3626-3630.
74. Pande VS, Grosberg AY, Tanaka T: **On the theory of folding kinetics for short proteins.** *Fold Des* 1997, **2**(2):109-114.
75. Doyle R, Simons K, Qian H, Baker D: **Local interactions and the optimization of protein folding.** *Proteins-Structure Function and Genetics* 1997, **29**(3):282-291.

76. Gross M: **Linguistic analysis of protein folding.** *FEBS Lett* 1996, **390**(3):249-252.
77. Unger R, Moult J: **Local interactions dominate folding in a simple protein model.** *J Mol Biol* 1996, **259**(5):988-994.
78. Fersht AR: **Mapping the Structures of Transition-States and Intermediates in Folding - Delineation of Pathways at High-Resolution.** *Philos T Roy Soc B* 1995, **348**(1323):11-15.
79. Govindarajan S, Goldstein RA: **Optimal Local Propensities for Model Proteins.** *Proteins-Structure Function and Genetics* 1995, **22**(4):413-418.
80. Orengo CA, Jones DT, Thornton JM: **Protein Superfamilies and Domain Superfolds.** *Nature* 1994, **372**(6507):631-634.
81. Dill KA, Fiebig KM, Chan HS: **Cooperativity in Protein-Folding Kinetics.** *Proc Natl Acad Sci* 1993, **90**(5):1942-1946.
82. Plaxco KW, Simons KT, Baker D: **Contact order, transition state placement and the refolding rates of single domain proteins.** *J Mol Biol* 1998, **277**(4):985-994.
83. van den Berg B, Ellis RJ, Dobson CM: **Effects of macromolecular crowding on protein folding and aggregation.** *Embo Journal* 1999, **18**(24):6927-6933.
84. Kim YE, Hipp MS, Bracher A, Hayer-Hartl M, Hartl FU: **Molecular Chaperone Functions in Protein Folding and Proteostasis.** *Annual Review of Biochemistry, Vol 82* 2013, **82**:323-355.
85. Zhang JZ: **Protein-length distributions for the three domains of life.** *Trends Genet* 2000, **16**(3):107-109.
86. Martin J, Hartl FU: **The effect of macromolecular crowding on chaperonin-mediated protein folding.** *Proc Natl Acad Sci U S A* 1997, **94**(4):1107-1112.
87. Balch WE, Morimoto RI, Dillin A, Kelly JW: **Adapting proteostasis for disease intervention.** *Science* 2008, **319**(5865):916-919.

88. Chiti F, Dobson CM: **Protein misfolding, functional amyloid, and human disease.** *Annu Rev Biochem* 2006, **75**:333-366.
89. Dill KA, Ozkan SB, Shell MS, Weikl TR: **The protein folding problem.** *Annu Rev Biophys* 2008, **37**:289-316.
90. Chou PY, Fasman GD: **Prediction of Protein Conformation.** *Biochemistry* 1974, **13**(2):222-245.
91. Chou PY, Fasman GD: **Empirical Predictions of Protein Conformation.** *Annu Rev Biochem* 1978, **47**:251-276.
92. Rost B, Eyrich VA: **EVA: Large-scale analysis of secondary structure prediction.** *Proteins-Structure Function and Genetics* 2001:192-199.
93. Li ZQ, Scheraga HA: **Monte-Carlo-Minimization Approach to the Multiple-Minima Problem in Protein Folding.** *Proc Natl Acad Sci* 1987, **84**(19):6611-6615.
94. Hansmann UHE, Okamoto Y: **Prediction of Peptide Conformation by Multicanonical Algorithm - New Approach to the Multiple-Minima Problem.** *J Comput Chem* 1993, **14**(11):1333-1338.
95. Sali A, Blundell TL: **Comparative Protein Modeling by Satisfaction of Spatial Restraints.** *J Mol Biol* 1993, **234**(3):779-815.
96. Jones DT, Taylor WR, Thornton JM: **A New Approach to Protein Fold Recognition.** *Nature* 1992, **358**(6381):86-89.
97. Moult J, Pedersen JT, Judson R, Fidelis K: **A Large-Scale Experiment to Assess Protein-Structure Prediction Methods.** *Proteins-Structure Function and Genetics* 1995, **23**(3):R2-R4.
98. Moult J: **A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction.** *Curr Opin Struct Biol* 2005, **15**(3):285-289.

99. Venclovas C, Zemla A, Fidelis K, Moult J: **Assessment of progress over the CASP experiments.** *Proteins-Structure Function and Bioinformatics* 2003, **53**:585-595.
100. Baker D: **Prediction and design of macromolecular structures and interactions.** *Philosophical Transactions of the Royal Society B-Biological Sciences* 2006, **361**(1467):459-463.
101. Bradley P, Misura KMS, Baker D: **Toward high-resolution de novo structure prediction for small proteins.** *Science* 2005, **309**(5742):1868-1871.
102. Zhang Y, Arakaki AK, Skolnick JR: **TASSER: An automated method for the prediction of protein tertiary structures in CASP6.** *Proteins-Structure Function and Bioinformatics* 2005, **61**:91-98.
103. Pieper U, Eswar N, Braberg H, Madhusudhan MS, Davis FP, Stuart AC, Mirkovic N, Rossi A, Marti-Renom MA, Fiser A *et al*: **MODBASE, a database of annotated comparative protein structure models, and associated resources.** *Nucleic Acids Res* 2004, **32**:D217-D222.
104. Tress M, Ezkurdia L, Grana O, Lopez G, Valencia A: **Assessment of predictions submitted for the CASP6 comparative modeling category.** *Proteins-Structure Function and Bioinformatics* 2005, **61**:27-45.
105. Duan Y, Kollman PA: **Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution.** *Science* 1998, **282**(5389):740-744.
106. Seibert MM, Patriksson A, Hess B, van der Spoel D: **Reproducible polypeptide folding and structure prediction using molecular dynamics simulations.** *J Mol Biol* 2005, **354**(1):173-183.
107. Freddolino PL, Liu F, Gruebele M, Schulten K: **Ten-microsecond molecular dynamics simulation of a fast-folding WW domain.** *Biophys J* 2008, **94**(10):L75-L77.
108. Piana S, Klepeis JL, Shaw DE: **Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations.** *Curr Opin Struct Biol* 2014, **24**:98-105.

109. Shaw DE: **Anton: A specialized machine for millisecond-scale molecular dynamics simulations of proteins.** *Abstr Pap Am Chem S* 2009, **238**.
110. Shaw DE, Dror RO, Salmon JK, Grossman JP, Mackenzie KM, Bank JA, Young C, Deneroff MM, Batson B, Bowers KJ *et al*: **Millisecond-Scale Molecular Dynamics Simulations on Anton.** *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis* 2009.
111. Piana S, Lindorff-Larsen K, Shaw DE: **Protein folding kinetics and thermodynamics from atomistic simulation.** *Proc Natl Acad Sci* 2012, **109**(44):17845-17850.
112. Piana S, Lindorff-Larsen K, Shaw DE: **Atomistic Description of the Folding of a Dimeric Protein.** *J Phys Chem B* 2013, **117**(42):12935-12942.
113. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, Shaw DE: **Improved side-chain torsion potentials for the Amber ff99SB protein force field.** *Proteins-Structure Function and Bioinformatics* 2010, **78**(8):1950-1958.
114. Best RB, Hummer G: **Optimized Molecular Dynamics Force Fields Applied to the Helix-Coil Transition of Polypeptides.** *J Phys Chem B* 2009, **113**(26):9004-9015.
115. Best RB, Zhu X, Shim J, Lopes PEM, Mittal J, Feig M, MacKerell AD: **Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone phi, psi and Side-Chain chi(1) and chi(2) Dihedral Angles.** *J Chem Theory Comput* 2012, **8**(9):3257-3273.
116. Li DW, Bruschweiler R: **NMR-Based Protein Potentials.** *Angew Chem Int Edit* 2010, **49**(38):6778-6780.
117. Piana S, Lindorff-Larsen K, Shaw DE: **How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization?** *Biophys J* 2011, **100**(9):L47-L49.
118. Braun AR, Sachs JN, Nagle JF: **Comparing Simulations of Lipid Bilayers to Scattering Data: The GROMOS 43A1-S3 Force Field.** *J Phys Chem B* 2013, **117**(17):5065-5072.

119. Mayne CG, Saam J, Schulten K, Tajkhorshid E, Gumbart JC: **Rapid Parameterization of Small Molecules Using the Force Field Toolkit.** *J Comput Chem* 2013, **34**(32):2757-2770.
120. Raval A, Piana S, Eastwood MP, Dror RO, Shaw DE: **Refinement of protein structure homology models via long, all-atom molecular dynamics simulations.** *Proteins-Structure Function and Bioinformatics* 2012, **80**(8):2071-2079.
121. Royer CA, Mann CJ, Matthews CR: **Resolution of the fluorescence equilibrium unfolding profile of trp aporepressor using single tryptophan mutants.** *Protein Sci* 1993, **2**(11):1844-1852.
122. Ladokhin AS, Jayasinghe S, White SH: **How to measure and analyze tryptophan fluorescence in membranes properly, and why bother?** *Anal Biochem* 2000, **285**(2):235-245.
123. Vivian JT, Callis PR: **Mechanisms of tryptophan fluorescence shifts in proteins.** *Biophys J* 2001, **80**(5):2093-2109.
124. Lakowicz JR: **Principles of fluorescence spectroscopy**, 3rd edn. New York: Springer; 2006.
125. Engelborghs Y: **The analysis of time resolved protein fluorescence in multi-tryptophan proteins.** *Spectrochim Acta A* 2001, **57**(11):2255-2270.
126. Matouschek A, Kellis JT, Jr., Serrano L, Fersht AR: **Mapping the transition state and pathway of protein folding by protein engineering.** *Nature* 1989, **340**(6229):122-126.
127. Sanchez IE, Kiefhaber T: **Hammond behavior versus ground state effects in protein folding: Evidence for narrow free energy barriers and residual structure in unfolded states.** *J Mol Biol* 2003, **327**(4):867-884.
128. Dalby PA, Oliveberg M, Fersht AR: **Folding intermediates of wild-type and mutants of barnase. I. Use of phi-value analysis and m-values to probe the cooperative nature of the folding pre-equilibrium.** *J Mol Biol* 1998, **276**(3):625-646.

129. Jackson SE, elMasry N, Fersht AR: **Structure of the hydrophobic core in the transition state for folding of chymotrypsin inhibitor 2: a critical test of the protein engineering method of analysis.** *Biochemistry* 1993, **32**(42):11270-11278.
130. Gromiha MM, Selvaraj S: **Important amino acid properties for determining the transition state structures of two-state protein mutants.** *FEBS Lett* 2002, **526**(1-3):129-134.
131. Northey JGB, Maxwell KL, Davidson AR: **Protein folding kinetics beyond the Phi value: Using multiple amino acid substitutions to investigate the structure of the SH3 domain folding transition state.** *J Mol Biol* 2002, **320**(2):389-402.
132. Sosnick TR, Dothager RS, Krantz BA: **Differences in the folding transition state of ubiquitin indicated by phi and psi analyses.** *Proc Natl Acad Sci* 2004, **101**(50):17377-17382.
133. Steward A, McDowell GS, Clarke J: **Topology is the Principal Determinant in the Folding of a Complex All-alpha Greek Key Death Domain from Human FADD.** *J Mol Biol* 2009, **389**(2):425-437.
134. Neuweiler H, Sharpe TD, Rutherford TJ, Johnson CM, Allen MD, Ferguson N, Fersht AR: **The Folding Mechanism of BBL: Plasticity of Transition-State Structure Observed within an Ultrafast Folding Protein Family.** *J Mol Biol* 2009, **390**(5):1060-1073.
135. Wensley BG, Gartner M, Choo WX, Batey S, Clarke J: **Different Members of a Simple Three-Helix Bundle Protein Family Have Very Different Folding Rate Constants and Fold by Different Mechanisms.** *J Mol Biol* 2009, **390**(5):1074-1085.
136. Fersht AR, Matouschek A, Serrano L: **The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding.** *J Mol Biol* 1992, **224**(3):771-782.
137. Zarrine-Afsar A, Davidson AR: **The analysis of protein folding kinetic data produced in protein engineering experiments.** *Methods* 2004, **34**(1):41-50.

138. Matouschek A, Fersht AR: **Protein engineering in analysis of protein folding pathways and stability.** *Methods Enzymol* 1991, **202**:82-112.
139. Kelly SM, Jess TJ, Price NC: **How to study proteins by circular dichroism.** *BBA* 2005, **1751**(2):119-139.
140. Kelly SM, Price NC: **Circular dichroism to study protein interactions.** *Curr Protoc Protein Sci* 2006, **Chapter 20**:Unit 20 10.
141. Zeeb M, Balbach J: **Protein folding studied by real-time NMR spectroscopy.** *Methods* 2004, **34**(1):65-74.
142. Zeeb M, Balbach J: **Millisecond protein folding studied by NMR spectroscopy.** *Protein Sci* 2004, **13**:168-168.
143. Fabian H, Naumann D: **Methods to study protein folding by stopped-flow FT-IR.** *Methods* 2004, **34**(1):28-40.
144. Kihara H, Semisotnov GV, Kotova NV, Kimura K, Amemiya Y, Serdyuk IN, Timchenko AA, Ikura T, Kuwajima K: **Kinetic study on protein folding studied by stopped-flow X-ray scattering.** *Prog Biophys Mol Bio* 1996, **65**:Pa332-Pa332.
145. Kuwajima K, Yamaya H, Sugai S: **The burst-phase intermediate in the refolding of beta-lactoglobulin studied by stopped-flow circular dichroism and absorption spectroscopy.** *J Mol Biol* 1996, **264**(4):806-822.
146. Kuwajima K: **The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure.** *Proteins* 1989, **6**(2):87-103.
147. Marqusee S, Robbins VH, Baldwin RL: **Unusually stable helix formation in short alanine-based peptides.** *Proc Natl Acad Sci U S A* 1989, **86**(14):5286-5290.
148. Roder H, Colon W: **Kinetic role of early intermediates in protein folding.** *Curr Opin Struct Biol* 1997, **7**(1):15-28.

149. Gruenewald B, Nicola CU, Lustig A, Schwarz G, Klump H: **Kinetics of the helix-coil transition of a polypeptide with non-ionic side groups, derived from ultrasonic relaxation measurements.** *Biophys Chem* 1979, **9**(2):137-147.
150. Shastry MC, Luck SD, Roder H: **A continuous-flow capillary mixing method to monitor reactions on the microsecond time scale.** *Biophys J* 1998, **74**(5):2714-2721.
151. Shastry MC, Roder H: **Evidence for barrier-limited protein folding kinetics on the microsecond time scale.** *Nat Struct Biol* 1998, **5**(5):385-392.
152. Xu M, Beresneva O, Rosario R, Roder H: **Microsecond Folding Dynamics of Apomyoglobin at Acidic pH.** *J Phys Chem B* 2012, **116**(23):7014-7025.
153. Bagshaw C: **Stopped-Flow Techniques.** In: *Encyclopedia of Biophysics*. Edited by Roberts GK: Springer Berlin Heidelberg; 2013: 2460-2466.
154. Brooks CL, 3rd: **Simulations of protein folding and unfolding.** *Curr Opin Struct Biol* 1998, **8**(2):222-226.
155. Chan HS, Dill KA: **Protein folding in the landscape perspective: chevron plots and non-Arrhenius kinetics.** *Proteins* 1998, **30**(1):2-33.
156. Dobson CM, Sali A, Karplus M: **Protein folding: a perspective from theory and experiment.** *Angew Chem Int Edit* 1998, **37**:868-893.
157. Jackson SE: **How do small single-domain proteins fold?** *Fold Des* 1998, **3**(4):R81-R91.
158. Whitmore L, Wallace BA: **Protein secondary structure analyses from circular dichroism spectroscopy: Methods and reference databases.** *Biopolymers* 2008, **89**(5):392-400.
159. Berova N, Nakanishi K, Woody R: **Circular dichroism : principles and applications**, 2nd edn. New York: Wiley-VCH; 2000.
160. Provencher SW, Glockner J: **Estimation of Globular Protein Secondary Structure from Circular-Dichroism.** *Biochemistry* 1981, **20**(1):33-37.

161. Manavalan P, Johnson WC: **Variable Selection Method Improves the Prediction of Protein Secondary Structure from Circular-Dichroism Spectra.** *Anal Biochem* 1987, **167**(1):76-85.
162. Andrade MA, Chacon P, Merelo JJ, Moran F: **Evaluation of Secondary Structure of Proteins from Uv Circular-Dichroism Spectra Using an Unsupervised Learning Neural-Network.** *Protein Eng* 1993, **6**(4):383-390.
163. Sreerama N, Woody RW: **A Self-Consistent Method for the Analysis of Protein Secondary Structure from Circular-Dichroism.** *Anal Biochem* 1993, **209**(1):32-44.
164. Johnson WC: **Analyzing protein circular dichroism spectra for accurate secondary structures.** *Proteins-Structure Function and Genetics* 1999, **35**(3):307-312.
165. Lobley A, Whitmore L, Wallace BA: **DICHROWEB: an interactive website for the analysis of protein secondary structure from circular dichroism spectra.** *Bioinformatics* 2002, **18**(1):211-212.
166. Whitmore L, Wallace BA: **DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data.** *Nucleic Acids Res* 2004, **32**(Web Server issue):W668-673.
167. Sreerama N, Woody RW: **Computation and analysis of protein circular dichroism spectra.** *Numerical Computer Methods, Pt D* 2004, **383**:318-351.
168. Sreerama N, Woody RW: **On the analysis of membrane protein circular dichroism spectra.** *Protein Sci* 2004, **13**(1):100-112.
169. Freskgard PO, Martensson LG, Jonasson P, Jonsson BH, Carlsson U: **Assignment of the Contribution of the Tryptophan Residues to the Circular-Dichroism Spectrum of Human Carbonic-Anhydrase .2.** *Biochemistry* 1994, **33**(47):14281-14288.
170. Woody AYM, Woody RW: **Individual tyrosine side-chain contributions to circular dichroism of ribonuclease.** *Biopolymers* 2003, **72**(6):500-513.

171. Boxer DH, Zhang H, Gourley DG, Hunter WN, Kelly SM, Price NC: **Sensing of remote oxyanion binding at the DNA binding domain of the molybdate-dependent transcriptional regulator, ModE.** *Org Biomol Chem* 2004, **2**(19):2829-2837.
172. Krell T, Horsburgh MJ, Cooper A, Kelly SM, Coggins JR: **Localization of the active site of type II dehydroquinases - Identification of a common arginine-containing motif in the two classes of dehydroquinases.** *J Biol Chem* 1996, **271**(40):24492-24497.
173. Hope J, Shearman MS, Baxter HC, Chong A, Kelly SM, Price NC: **Cytotoxicity of prion protein peptide (PrP106-126) differs in mechanism from the cytotoxic activity of the Alzheimer's disease amyloid peptide, A beta 25-35.** *Neurodegeneration* 1996, **5**(1):1-11.
174. Roberts GCK: **NMR of macromolecules : a practical approach.** Oxford ; New York: IRL Press at Oxford University Press; 1993.
175. Takegoshi K, Nakamura S, Terao T: **^{13}C - ^1H dipolar-driven ^{13}C - ^{13}C recoupling without ^{13}C rf irradiation in nuclear magnetic resonance of rotating solids.** *Journal of Chemical Physics* 2003, **118**:2325-2341.
176. Takegoshi K, Nakamura S, Terao T: **^{13}C - ^1H dipolar-assisted rotational resonance in magic-angle spinning NMR.** *Chem Phys Lett* 2001, **344**(5):631-637.
177. Uversky VN: **Amyloidogenesis of natively unfolded proteins.** *Curr Alzheimer Res* 2008, **5**(3):260-287.
178. Fandrich M: **On the structural definition of amyloid fibrils and other polypeptide aggregates.** *Cell Mol Life Sci* 2007, **64**(16):2066-2078.
179. Hortschansky P, Christopeit T, Schroeckh V, Fandrich M: **Thermodynamic analysis of the aggregation propensity of oxidized Alzheimer's beta-amyloid variants.** *Protein Sci* 2005, **14**(11):2915-2918.
180. Uversky VN, Fink AL: **Conformational constraints for amyloid fibrillation: the importance of being unfolded.** *Biochim Biophys Acta* 2004, **1698**(2):131-153.

181. Chiti F, Webster P, Taddei N, Clark A, Stefani M, Ramponi G, Dobson CM: **Designing conditions for in vitro formation of amyloid protofilaments and fibrils.** *Proc Natl Acad Sci* 1999, **96**(7):3590-3594.
182. Fandrich M, Forge V, Buder K, Kittler M, Dobson CM, Diekmann S: **Myoglobin forms amyloid fibrils by association of unfolded polypeptide segments.** *Proc Natl Acad Sci* 2003, **100**(26):15463-15468.
183. Collins JC, Greene LH: **Biophysical Analysis of the Transition of an All alpha-Helical Greek-Key Protein into Amyloid Fibrils Composed of beta-Sheet Structure.** *Protein Pept Lett* 2012, **19**(9):982-990.
184. Scheidt HA, Morgado I, Rothmund S, Huster D, Fandrich M: **Solid-State NMR Spectroscopic Investigation of A beta Protofibrils: Implication of a beta-Sheet Remodeling upon Maturation into Terminal Amyloid Fibrils.** *Angew Chem Int Edit* 2011, **50**(12):2837-2840.
185. Giurleo JT, He XL, Talaga DS: **beta-lactoglobulin assembles into amyloid through sequential aggregated intermediates.** *J Mol Biol* 2008, **381**(5):1332-1348.
186. Harper JD, Lansbury PT, Jr.: **Models of amyloid seeding in Alzheimer's disease and scrapie: mechanistic truths and physiological consequences of the time-dependent solubility of amyloid proteins.** *Annu Rev Biochem* 1997, **66**:385-407.
187. Uversky VN, Lyubchenko YL: **Bio-nanoimaging : Protein Misfolding & Aggregation**, vol. 1, 1st edn: Elsevier.
188. Bemporad F, Calloni G, Campioni S, Plakoutsi G, Taddei N, Chiti F: **Sequence and structural determinants of amyloid fibril formation.** *Acc Chem Res* 2006, **39**(9):620-627.
189. Greenwald J, Riek R: **Biology of amyloid: structure, function, and regulation.** *Structure* 2010, **18**(10):1244-1260.
190. Lu JX, Qiang W, Yau WM, Schwieters CD, Meredith SC, Tycko R: **Molecular Structure of beta-Amyloid Fibrils in Alzheimer's Disease Brain Tissue.** *Cell* 2013, **154**(6):1257-1268.

191. Jahn TR, Radford SE: **Folding versus aggregation: Polypeptide conformations on competing pathways.** *Arch Biochem Biophys* 2008, **469**(1):100-117.
192. Colletier JP, Laganowsky A, Landau M, Zhao ML, Soriaga AB, Goldschmidt L, Flot D, Cascio D, Sawaya MR, Eisenberg D: **Molecular basis for amyloid-beta polymorphism.** *Proc Natl Acad Sci* 2011, **108**(41):16938-16943.
193. Makin OS, Atkins E, Sikorski P, Johansson J, Serpell LC: **Molecular basis for amyloid fibril formation and stability.** *Proc Natl Acad Sci* 2005, **102**(2):315-320.
194. Huff ME, Balch WE, Kelly JW: **Pathological and functional amyloid formation orchestrated by the secretory pathway.** *Curr Opin Struct Biol* 2003, **13**(6):674-682.
195. Maji SK, Perrin MH, Sawaya MR, Jessberger S, Vadodaria K, Rissman RA, Singru PS, Nilsson KPR, Simon R, Schubert D *et al*: **Functional Amyloids As Natural Storage of Peptide Hormones in Pituitary Secretory Granules.** *Science* 2009, **325**(5938):328-332.
196. Bennett MJ, Sawaya MR, Eisenberg D: **Deposition diseases and 3D domain swapping.** *Structure* 2006, **14**(5):811-824.
197. Eisenberg D, Nelson R, Sawaya MR, Balbirnie M, Sambashivan S, Ivanova MI, Madsen AO, Riek C: **The structural biology of protein aggregation diseases: Fundamental questions and some answers.** *Acc Chem Res* 2006, **39**(9):568-575.
198. Nelson R, Eisenberg D: **Recent atomic models of amyloid fibril structure.** *Curr Opin Struct Biol* 2006, **16**(2):260-265.
199. Nelson R, Eisenberg D: **Structural models of amyloid-like fibrils.** *Adv Protein Chem* 2006, **73**:235-282.
200. Osherovich LZ, Weissman JS: **The utility of prions.** *Dev Cell* 2002, **2**(2):143-151.
201. True HL, Lindquist SL: **A yeast prion provides a mechanism for genetic variation and phenotypic diversity.** *Nature* 2000, **407**(6803):477-483.

202. Maddelein ML, Dos Reis S, Duvezin-Caubet S, Coulary-Salin B, Saupe SJ: **Amyloid aggregates of the HET-s prion protein are infectious.** *Proc Natl Acad Sci USA* 2002, **99**(11):7402-7407.
203. Dobson CM: **Principles of protein folding, misfolding and aggregation.** *Semin Cell Dev Biol* 2004, **15**(1):3-16.
204. Chiti F, Taddei N, Baroni F, Capanni C, Stefani M, Ramponi G, Dobson CM: **Kinetic partitioning of protein folding and aggregation.** *Nat Struct Biol* 2002, **9**(2):137-143.
205. Lopez de la Paz M, Serrano L: **Sequence determinants of amyloid fibril formation.** *Proc Natl Acad Sci U S A* 2004, **101**(1):87-92.
206. Lindorff-Larsen K, Rogen P, Paci E, Vendruscolo M, Dobson CM: **Protein folding and the organization of the protein topology universe.** *Trends Biochem Sci* 2005, **30**(1):13-19.
207. Esteras-Chopo A, Serrano L, Lopez de la Paz M: **The amyloid stretch hypothesis: recruiting proteins toward the dark side.** *Proc Natl Acad Sci U S A* 2005, **102**(46):16672-16677.
208. Chiti F, Stefani M, Taddei N, Ramponi G, Dobson CM: **Rationalization of the effects of mutations on peptide and protein aggregation rates.** *Nature* 2003, **424**(6950):805-808.
209. DuBay KF, Pawar AP, Chiti F, Zurdo J, Dobson CM, Vendruscolo M: **Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains.** *J Mol Biol* 2004, **341**(5):1317-1326.
210. Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L: **Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins.** *Nat Biotechnol* 2004, **22**(10):1302-1306.
211. Maurer-Stroh S, Debulpaep M, Kuemmerer N, Lopez de la Paz M, Martins IC, Reumers J, Morris KL, Copland A, Serpell L, Serrano L *et al*: **Exploring the sequence determinants of amyloid structure using position-specific scoring matrices.** *Nat Methods* 2010, **7**(3):237-242.

212. Tartaglia GG, Cavalli A, Pellarin R, Caflisch A: **Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences.** *Protein Sci* 2005, **14**(10):2723-2734.
213. Rousseau F, Serrano L, Schymkowitz JWH: **How evolutionary pressure against protein aggregation shaped chaperone specificity.** *J Mol Biol* 2006, **355**(5):1037-1047.
214. Linding R, Schymkowitz J, Rousseau F, Diella F, Serrano L: **A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins.** *J Mol Biol* 2004, **342**(1):345-353.
215. Wigley WC, Corboy MJ, Cutler TD, Thibodeau PH, Oldan J, Lee MG, Rizo J, Hunt JF, Thomas PJ: **A protein sequence that can encode native structure by disfavoring alternate conformations.** *Nat Struct Biol* 2002, **9**(5):381-388.
216. Richardson JS, Richardson DC: **Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation.** *Proc Natl Acad Sci* 2002, **99**(5):2754-2759.
217. Parrini C, Taddei N, Ramazzotti M, Degl'Innocenti D, Ramponi G, Dobson CM, Chiti F: **Glycine residues appear to be evolutionarily conserved for their ability to inhibit aggregation.** *Structure* 2005, **13**(8):1143-1151.
218. Jahn TR, Parker MJ, Homans SW, Radford SE: **Amyloid formation under physiological conditions proceeds via a native-like folding intermediate.** *Nat Struct Mol Biol* 2006, **13**(3):195-201.
219. Liu K, Cho HS, Hoyt DW, Nguyen TN, Olds P, Kelly JW, Wemmer DE: **Deuterium-proton exchange on the native wild-type transthyretin tetramer identifies the stable core of the individual subunits and indicates mobility at the subunit interface.** *J Mol Biol* 2000, **303**(4):555-565.
220. Nordlund A, Oliveberg M: **Folding of Cu/Zn superoxide dismutase suggests structural hotspots for gain of neurotoxic function in ALS: Parallels to precursors in amyloid disease.** *Proc Natl Acad Sci* 2006, **103**(27):10218-10223.

221. Steward A, Adhya S, Clarke J: **Sequence conservation in Ig-like domains: the role of highly conserved proline residues in the fibronectin type III superfamily.** *J Mol Biol* 2002, **318**(4):935-940.
222. Drueke TB, Massy ZA: **Beta2-microglobulin.** *Semin Dial* 2009, **22**(4):378-380.
223. Jaroniec CP, MacPhee CE, Astrof NS, Dobson CM, Griffin RG: **Molecular conformation of a peptide fragment of transthyretin in an amyloid fibril.** *Proc Natl Acad Sci U S A* 2002, **99**(26):16748-16753.
224. Mak CM, Kwong YL, Lam CW, Chan SC, Lo CM, Fan ST, Chang CM, Lau YK, Lok-Sun U, Tam S: **Identification of a novel TTR Gly67Glu mutant and the first case series of familial transthyretin amyloidosis in Hong Kong Chinese.** *Amyloid-Journal of Protein Folding Disorders* 2007, **14**(4):293-297.
225. Museth AK, Brorsson AC, Lundqvist M, Tibell LAE, Jonsson BH: **The ALS-Associated Mutation G93A in Human Copper-Zinc Superoxide Dismutase Selectively Destabilizes the Remote Metal Binding Region.** *Biochemistry* 2009, **48**(37):8817-8829.
226. Lindberg MJ, Tibell L, Oliveberg M: **Common denominator of Cu/Zn superoxide dismutase mutants associated with amyotrophic lateral sclerosis: Decreased stability of the apo state.** *Proc Natl Acad Sci* 2002, **99**(26):16607-16612.
227. Broome BM, Hecht MH: **Nature disfavors sequences of alternating polar and non-polar amino acids: Implications for amyloidogenesis.** *J Mol Biol* 2000, **296**(4):961-968.
228. Wisniewski T, Golabek AA, Kida E, Wisniewski KE, Frangione B: **Conformational Mimicry in Alzheimers-Disease - Role of Apolipoproteins in Amyloidogenesis.** *Am J Pathol* 1995, **147**(2):238-244.
229. Jansen R, Dzwolak W, Winter R: **Amyloidogenic self-assembly of insulin aggregates probed by high resolution atomic force microscopy.** *Biophys J* 2005, **88**(2):1344-1353.
230. Gustafsson M, Thyberg J, Naslund J, Eliasson E, Johansson J: **Amyloid fibril formation by pulmonary surfactant protein C.** *FEBS Lett* 1999, **464**(3):138-142.

231. Dannies PS: **Mechanisms for storage of prolactin and growth hormone in secretory granules.** *Mol Genet Metab* 2002, **76**(1):6-13.
232. Rao PN, Reddy KS, Bhuyan AK: **Amyloid fibrillation of human Apaf-1 CARD.** *Biochemistry* 2009, **48**(32):7656-7664.
233. de Groot NS, Ventura S: **Amyloid fibril formation by bovine cytochrome c.** *Spectros-Int J* 2005, **19**(4):199-205.
234. Holm NK, Jespersen SK, Thomassen LV, Wolff TY, Sehgal P, Thomsen LA, Christiansen G, Andersen CB, Knudsen AD, Otzen DE: **Aggregation and fibrillation of bovine serum albumin.** *BBA* 2007, **1774**(9):1128-1138.
235. Vetri V, D'Amico M, Fodera V, Leone M, Ponzoni A, Sberveglieri G, Militello V: **Bovine Serum Albumin protofibril-like aggregates formation: Solo but not simple mechanism.** *Arch Biochem Biophys* 2011, **508**(1):13-24.
236. Pertinhez TA, Bouchard M, Tomlinson EJ, Wain R, Ferguson SJ, Dobson CM, Smith LJ: **Amyloid fibril formation by a helical cytochrome.** *FEBS Lett* 2001, **495**(3):184-186.
237. Hung YT, Lin MS, Chen WY, Wang SS: **Investigating the effects of sodium dodecyl sulfate on the aggregative behavior of hen egg-white lysozyme at acidic pH.** *Colloids Surf B, Biointerfaces* 2010, **81**(1):141-151.
238. Johnson KH, Sletten K, Hayden DW, O'Brien TD, Roertgen KE, Westermark P: **Pulmonary vascular amyloidosis in aged dogs. A new form of spontaneously occurring amyloidosis derived from apolipoprotein AI.** *Am J Pathol* 1992, **141**(5):1013-1019.
239. Westermark P, Mucchiano G, Marthin T, Johnson KH, Sletten K: **Apolipoprotein A1-derived amyloid in human aortic atherosclerotic plaques.** *Am J Pathol* 1995, **147**(5):1186-1192.
240. Amarzguioui M, Mucchiano G, Haggqvist B, Westermark P, Kavlie A, Sletten K, Prydz H: **Extensive intimal apolipoprotein A1-derived amyloid deposits in a patient with an apolipoprotein A1 mutation.** *Biochem Biophys Res Commun* 1998, **242**(3):534-539.

241. Mucchiano GI, Haggqvist B, Sletten K, Westermark P: **Apolipoprotein A-I-derived amyloid in atherosclerotic plaques of the human aorta.** *J Pathol* 2001, **193**(2):270-275.
242. Tall AR: **Cholesterol efflux pathways and other potential mechanisms involved in the athero-protective effect of high density lipoproteins.** *J Intern Med* 2008, **263**(3):256-273.
243. Rye KA, Barter PJ: **Formation and metabolism of prebeta-migrating, lipid-poor apolipoprotein A-I.** *Arterioscler Thromb Vasc Biol* 2004, **24**(3):421-428.
244. Soutar AK, Hawkins PN, Vigushin DM, Tennent GA, Booth SE, Hutton T, Nguyen O, Totty NF, Feest TG, Hsuan JJ *et al*: **Apolipoprotein AI mutation Arg-60 causes autosomal dominant amyloidosis.** *Proc Natl Acad Sci* 1992, **89**(16):7389-7393.
245. Vigushin DM, Gough J, Allan D, Alguacil A, Penner B, Pettigrew NM, Quinonez G, Bernstein K, Booth SE, Booth DR *et al*: **Familial nephropathic systemic amyloidosis caused by apolipoprotein AI variant Arg26.** *Q J Med* 1994, **87**(3):149-154.
246. Booth DR, Tan SY, Booth SE, Hsuan JJ, Totty NF, Nguyen O, Hutton T, Vigushin DM, Tennent GA, Hutchinson WL *et al*: **A new apolipoprotein AI variant, Trp50Arg, causes hereditary amyloidosis.** *Q J Med* 1995, **88**(10):695-702.
247. Booth DR, Tan SY, Booth SE, Tennent GA, Hutchinson WL, Hsuan JJ, Totty NF, Truong O, Soutar AK, Hawkins PN *et al*: **Hereditary hepatic and systemic amyloidosis caused by a new deletion/insertion mutation in the apolipoprotein AI gene.** *J Clin Invest* 1996, **97**(12):2714-2721.
248. Hamidi Asl L, Liepnieks JJ, Hamidi Asl K, Uemichi T, Moulin G, Desjoyaux E, Loire R, Delpech M, Grateau G, Benson MD: **Hereditary amyloid cardiomyopathy caused by a variant apolipoprotein A1.** *Am J Pathol* 1999, **154**(1):221-227.
249. Hamidi Asl K, Liepnieks JJ, Nakamura M, Parker F, Benson MD: **A novel apolipoprotein A-1 variant, Arg173Pro, associated with cardiac and cutaneous amyloidosis.** *Biochem Biophys Res Commun* 1999, **257**(2):584-588.

250. Wong YQ, Binger KJ, Howlett GJ, Griffin MD: **Identification of an amyloid fibril forming peptide comprising residues 46-59 of apolipoprotein A-I.** *FEBS Lett* 2012, **586**(13):1754-1758.
251. Gursky O, Atkinson D: **Thermal unfolding of human high-density apolipoprotein A-1: implications for a lipid-free molten globular state.** *Proc Natl Acad Sci* 1996, **93**(7):2991-2995.
252. Zehender F, Ziegler A, Schonfeld HJ, Seelig J: **Thermodynamics of protein self-association and unfolding. The case of apolipoprotein A-I.** *Biochemistry* 2012, **51**(6):1269-1280.
253. Wong YQ, Binger KJ, Howlett GJ, Griffin MD: **Methionine oxidation induces amyloid fibril formation by full-length apolipoprotein A-I.** *Proc Natl Acad Sci* 2010, **107**(5):1977-1982.
254. Blundell TL, Cutfield JF, Cutfield SM, Dodson EJ, Dodson GG, Hodgkin DC, Mercola DA: **Three-dimensional atomic structure of insulin and its relationship to activity.** *Diabetes* 1972, **21**(2 Suppl):492-505.
255. Blundell TL, Cutfield JF, Dodson EJ, Dodson GG, Hodgkin DC, Mercola DA: **The crystal structure of rhombohedral 2 zinc insulin.** *Cold Spring Harb Symp Quant Biol* 1972, **36**:233-241.
256. Brange J, Andersen L, Laursen ED, Meyn G, Rasmussen E: **Toward understanding insulin fibrillation.** *J Pharm Sci* 1997, **86**(5):517-525.
257. Bryant C, Spencer DB, Miller A, Bakaysa DL, McCune KS, Maple SR, Pekar AH, Brems DN: **Acid stabilization of insulin.** *Biochemistry* 1993, **32**(32):8075-8082.
258. Westermark P, Wilander E: **Islet Amyloid in Type-2 (Non-Insulin-Dependent) Diabetes Is Related to Insulin.** *Diabetologia* 1983, **24**(5):342-346.
259. Westermark P, Wernstedt C, Wilander E, Hayden DW, O'Brien TD, Johnson KH: **Amyloid Fibrils in Human Insulinoma and Islets of Langerhans of the Diabetic Cat Are Derived from a Neuropeptide-Like Protein Also Present in Normal Islet Cells.** *Proc Natl Acad Sci* 1987, **84**(11):3881-3885.

260. Ehrlich JC, Ratner IM: **Amyloidosis of Islets of Langerhans - a Restudy of Islet Hyalin in Diabetic and Nondiabetic Individuals.** *Am J Pathol* 1961, **38**(1):49-&.
261. Nielsen L, Khurana R, Coats A, Frokjaer S, Brange J, Vyas S, Uversky VN, Fink AL: **Effect of environmental factors on the kinetics of insulin fibril formation: elucidation of the molecular mechanism.** *Biochemistry* 2001, **40**(20):6036-6046.
262. Sluzky V, Tamada JA, Klivanov AM, Langer R: **Kinetics of insulin aggregation in aqueous solutions upon agitation in the presence of hydrophobic surfaces.** *Proc Natl Acad Sci* 1991, **88**(21):9377-9381.
263. Nielsen L, Frokjaer S, Brange J, Uversky VN, Fink AL: **Probing the mechanism of insulin fibril formation with insulin mutants.** *Biochemistry* 2001, **40**(28):8397-8409.
264. Szyperski T, Vandenbussche G, Curstedt T, Ruyschaert JM, Wuthrich K, Johansson J: **Pulmonary surfactant-associated polypeptide C in a mixed organic solvent transforms from a monomeric alpha-helical state into insoluble beta-sheet aggregates.** *Protein Sci* 1998, **7**(12):2533-2540.
265. Johansson J: **Structure and properties of surfactant protein C.** *BBA* 1998, **1408**(2-3):161-172.
266. Johansson J, Nilsson G, Stromberg R, Robertson B, Jornvall H, Curstedt T: **Secondary structure and biophysical activity of synthetic analogues of the pulmonary surfactant polypeptide SP-C.** *Biochem J* 1995, **307** (Pt 2):535-541.
267. Freeman ME, Kanyicska B, Lerant A, Nagy G: **Prolactin: structure, function, and regulation of secretion.** *Physiol Rev* 2000, **80**(4):1523-1631.
268. Caroppi P, Sinibaldi F, Fiorucci L, Santucci R: **Apoptosis and human diseases: mitochondrion damage and lethal role of released cytochrome C as proapoptotic protein.** *Curr Med Chem* 2009, **16**(31):4058-4065.
269. Zou H, Li Y, Liu X, Wang X: **An APAF-1.cytochrome c multimeric complex is a functional apoptosome that activates procaspase-9.** *J Biol Chem* 1999, **274**(17):11549-11556.

270. Bratton SB, Salvesen GS: **Regulation of the Apaf-1-caspase-9 apoptosome.** *J Cell Sci* 2010, **123**(Pt 19):3209-3214.
271. Ivins KJ, Thornton PL, Rohn TT, Cotman CW: **Neuronal apoptosis induced by beta-amyloid is mediated by caspase-8.** *Neurobiol Dis* 1999, **6**(5):440-449.
272. Su JH, Anderson AJ, Cribbs DH, Tu C, Tong L, Kesslack P, Cotman CW: **Fas and Fas ligand are associated with neuritic degeneration in the AD brain and participate in beta-amyloid-induced neuronal death.** *Neurobiol Dis* 2003, **12**(3):182-193.
273. Kanga C, Krishnamurthy S, Shiva S: **Myoglobin and mitochondria: a relationship bound by oxygen and nitric oxide.** *Nitric Oxide* 2012, **26**(4):251-258.
274. Rassaf T, Flogel U, Drexhage C, Hendgen-Cotta U, Kelm M, Schrader J: **Nitrite reductase function of deoxymyoglobin - Oxygen sensor and regulator of cardiac energetics and function.** *Circ Res* 2007, **100**(12):1749-1754.
275. Shiva S, Huang Z, Grubina R, Sun JH, Ringwood LA, MacArthur PH, Xu XL, Murphy E, Darley-Usmar VM, Gladwin MT: **Deoxymyoglobin is a nitrite reductase that generates nitric oxide and regulates mitochondrial respiration.** *Circ Res* 2007, **100**(5):654-661.
276. Fandrich M, Fletcher MA, Dobson CM: **Amyloid fibrils from muscle myoglobin - Even an ordinary globular protein can assume a rogue guise if conditions are right.** *Nature* 2001, **410**(6825):165-166.
277. Sirangelo I, Malmo C, Iannuzzi C, Mezzogiorno A, Bianco MR, Papa M, Irace G: **Fibrillogenesis and cytotoxic activity of the amyloid-forming apomyoglobin mutant W7FW14F.** *J Biol Chem* 2004, **279**(13):13183-13189.
278. Sirangelo I, Malmo C, Casillo M, Mezzogiorno A, Papa M, Irace G: **Tryptophanyl substitutions in apomyoglobin determine protein aggregation and amyloid-like fibril formation at physiological pH.** *J Biol Chem* 2002, **277**(48):45887-45891.
279. Katina NS, Ilyina NB, Kashparov IA, Balobanov VA, Vasiliev VD, Bychkova VE: **Apomyoglobin mutants with single point mutations at Val10 can form**

- amyloid structures at permissive temperature. *Biochemistry-Moscow* 2011, 76(5):555-563.**
280. Roche M, Rondeau P, Singh NR, Tarnus E, Bourdon E: **The antioxidant properties of serum albumin. *FEBS Lett* 2008, 582(13):1783-1787.**
 281. Bhattacharya M, Jain N, Mukhopadhyay S: **Insights into the Mechanism of Aggregation and Fibril Formation from Bovine Serum Albumin. *J Phys Chem B* 2011, 115(14):4195-4205.**
 282. Frare E, Mossuto MF, Polverino de Laureto P, Dumoulin M, Dobson CM, Fontana A: **Identification of the core structure of lysozyme amyloid fibrils by proteolysis. *J Mol Biol* 2006, 361(3):551-561.**
 283. Kheterpal I, Williams A, Murphy C, Bledsoe B, Wetzel R: **Structural features of the A beta amyloid fibril elucidated by limited proteolysis. *Biochemistry* 2001, 40(39):11757-11767.**
 284. Monti M, Amoresano A, Giorgetti S, Bellotti V, Pucci P: **Limited proteolysis in the investigation of beta 2-microglobulin amyloidogenic and fibrillar states. *BBA - Proteins Proteom* 2005, 1753(1):44-50.**
 285. Militello V, Casarino C, Emanuele A, Giostra A, Pullara F, Leone M: **Aggregation kinetics of bovine serum albumin studied by FTIR spectroscopy and light scattering. *Biophys Chem* 2004, 107(2):175-187.**
 286. Kragh-Hansen U: **Structure and ligand binding properties of human serum albumin. *Dan Med Bull* 1990, 37(1):57-84.**
 287. Haass C, Selkoe DJ: **Soluble protein oligomers in neurodegeneration: lessons from the Alzheimer's amyloid beta-peptide. *Nat Rev Mol Cell Bio* 2007, 8(2):101-112.**
 288. Groenning M: **Binding mode of Thioflavin T and other molecular probes in the context of amyloid fibrils-current status. *J Chem Biol* 2010, 3(1):1-18.**
 289. Freire S, de Araujo MH, Al-Soufi W, Novo M: **Photophysical study of Thioflavin T as fluorescence marker of amyloid fibrils. *Dyes Pigments* 2014, 110:97-105.**

290. Vassar PS, Culling CFA: **Fluorescent Stains, with Special Reference to Amyloid and Connective Tissues.** *Arch Pathol* 1959, **68**(5):487-498.
291. Saeed SM, Fine G: **Thioflavin-T for Amyloid Detection.** *Am J Clin Pathol* 1967, **47**(5):588-&.
292. Ban T, Hamada D, Hasegawa K, Naiki H, Goto Y: **Direct observation of amyloid fibril growth monitored by thioflavin T fluorescence.** *J Biol Chem* 2003, **278**(19):16462-16465.
293. Andersen CB, Yagi H, Manno M, Martorana V, Ban T, Christiansen G, Otzen DE, Goto Y, Rischel C: **Branching in Amyloid Fibril Growth.** *Biophys J* 2009, **96**(4):1529-1536.
294. Sabate R, Saupe SJ: **Thioflavin T fluorescence anisotropy: an alternative technique for the study of amyloid aggregation.** *Biochem Biophys Res Commun* 2007, **360**(1):135-138.
295. Naiki H, Higuchi K, Nakakuki K, Takeda T: **Kinetic-Analysis of Amyloid Fibril Polymerization Invitro.** *Laboratory Investigation* 1991, **65**(1):104-110.
296. LeVine H, 3rd: **Stopped-flow kinetics reveal multiple phases of thioflavin T binding to Alzheimer beta (1-40) amyloid fibrils.** *Arch Biochem Biophys* 1997, **342**(2):306-316.
297. LeVine H, 3rd: **Thioflavine T interaction with synthetic Alzheimer's disease beta-amyloid peptides: detection of amyloid aggregation in solution.** *Protein Sci* 1993, **2**(3):404-410.
298. Pedersen JS, Dikov D, Flink JL, Hjuler HA, Christiansen G, Otzen DE: **The changing face of glucagon fibrillation: Structural polymorphism and conformational imprinting.** *J Mol Biol* 2006, **355**(3):501-523.
299. Wood SJ, Maleeff B, Hart T, Wetzel R: **Physical, morphological and functional differences between pH 5.8 and 7.4 aggregates of the Alzheimer's amyloid peptide AP.** *J Mol Biol* 1996, **256**(5):870-877.
300. Kardos J, Okuno D, Kawai T, Hagihara Y, Yumoto N, Kitagawa T, Zavodszky P, Naiki H, Goto Y: **Structural studies reveal that the diverse morphology of**

- beta(2)-microglobulin aggregates is a reflection of different molecular architectures.** *Biochim Biophys Acta* 2005, **1753**(1):108-120.
301. Wall J, Murphy CL, Solomon A: **In vitro immunoglobulin light chain fibrillogenesis.** *Methods Enzymol* 1999, **309**:204-217.
 302. Ahn JS, Lee JH, Kim JH, Paik SR: **Novel method for quantitative determination of amyloid fibrils of alpha-synuclein and amyloid beta/A4 protein by using resveratrol.** *Anal Biochem* 2007, **367**(2):259-265.
 303. D'Amico M, Di Carlo MG, Groenning M, Militello V, Vetri V, Leone M: **Thioflavin T Promotes A beta(1-40) Amyloid Fibrils Formation.** *J Phys Chem Lett* 2012, **3**(12):1596-1601.
 304. Reinke AA, Abulwerdi GA, Gestwicki JE: **Quantifying Prefibrillar Amyloids in vitro by Using a "Thioflavin-Like" Spectroscopic Method.** *Chembiochem* 2010, **11**(13):1889-1895.
 305. Selkoe DJ: **Folding proteins in fatal ways.** *Nature* 2003, **426**(6968):900-904.
 306. Gong YS, Chang L, Viola KL, Lacor PN, Lambert MP, Finch CE, Krafft GA, Klein WL: **Alzheimer's disease-affected brain: Presence of oligomeric A beta ligands (ADDLs) suggests a molecular basis for reversible memory loss.** *Proc Natl Acad Sci* 2003, **100**(18):10417-10422.
 307. Giorgadze TA, Shiina N, Baloch ZW, Tomaszewski JE, Gupta PK: **Improved detection of amyloid in fat pad aspiration: an evaluation of Congo red stain by fluorescent microscopy.** *Diagn Cytopathol* 2004, **31**(5):300-306.
 308. Sen S, Basdemir G: **Diagnosis of renal amyloidosis using Congo red fluorescence.** *Pathol Int* 2003, **53**(8):534-538.
 309. Howie AJ, Brewer DB: **Optical properties of amyloid stained by Congo red: History and mechanisms.** *Micron* 2009, **40**(3):285-301.
 310. McParland VJ, Kad NM, Kalverda AP, Brown A, Kirwin-Jones P, Hunter MG, Sunde M, Radford SE: **Partially unfolded states of beta(2)-microglobulin and amyloid formation in vitro.** *Biochemistry* 2000, **39**(30):8735-8746.

311. Turnell WG, Finch JT: **Binding of the Dye Congo Red to the Amyloid Protein Pig Insulin Reveals a Novel Homology Amongst Amyloid-Forming Peptide Sequences.** *J Mol Biol* 1992, **227**(4):1205-1223.
312. Kim YS, Randolph TW, Manning MC, Stevens FJ, Carpenter JF: **Congo red populates partially unfolded states of an amyloidogenic protein to enhance aggregation and amyloid fibril formation.** *J Biol Chem* 2003, **278**(12):10842-10850.
313. Klunk WE, Jacob RF, Mason RP: **Quantifying amyloid beta-peptide (Abeta) aggregation using the Congo red-Abeta (CR-abeta) spectrophotometric assay.** *Anal Biochem* 1999, **266**(1):66-76.
314. Klunk WE, Pettegrew JW, Abraham DJ: **Quantitative evaluation of congo red binding to amyloid-like proteins with a beta-pleated sheet conformation.** *J Histochem Cytochem* 1989, **37**(8):1273-1281.
315. Porat Y, Abramowitz A, Gazit E: **Inhibition of amyloid fibril formation by polyphenols: Structural similarity and aromatic interactions as a common inhibition mechanism.** *Chem Biol Drug Des* 2006, **67**(1):27-37.
316. Caughey B, Ernst D, Race RE: **Congo red inhibition of scrapie agent replication.** *J Virol* 1993, **67**(10):6270-6272.
317. Lorenzo A, Yankner BA: **Beta-amyloid neurotoxicity requires fibril formation and is inhibited by congo red.** *Proc Natl Acad Sci U S A* 1994, **91**(25):12243-12247.
318. Chander H, Chauhan A, Chauhan V: **Binding of proteases to fibrillar amyloid-beta protein and its inhibition by Congo red.** *J Alzheimers Dis* 2007, **12**(3):261-269.
319. Thompson LDR, Derringer GA, Wenig BM: **Amyloidosis of the larynx: A clinicopathologic study of 11 cases.** *Modern Pathol* 2000, **13**(5):528-535.
320. Krebs MRH, Bromley EHC, Donald AM: **The binding of thioflavin-T to amyloid fibrils: localisation and implications.** *J Struct Biol* 2005, **149**(1):30-37.

321. Stsiapura VI, Maskevich AA, Kuzmitsky VA, Uversky VN, Kuznetsova IM, Turoverov KK: **Thioflavin T as a Molecular Rotor: Fluorescent Properties of Thioflavin T in Solvents with Different Viscosity.** *J Phys Chem B* 2008, **112**(49):15893-15902.
322. Stsiapura VI, Maskevich AA, Kuzmitsky VA, Turoverov KK, Kuznetsova IM: **Computational study of thioflavin T torsional relaxation in the excited state.** *Journal of Physical Chemistry A* 2007, **111**(22):4829-4835.
323. Friedhoff P, Schneider A, Mandelkow EM, Mandelkow E: **Rapid assembly of Alzheimer-like paired helical filaments from microtubule-associated protein tau monitored by fluorescence in solution.** *Biochemistry* 1998, **37**(28):10223-10230.
324. Wu C, Wang Z, Lei H, Zhang W, Duan Y: **Dual binding modes of Congo red to amyloid protofibril surface observed in molecular dynamics simulations.** *J Am Chem Soc* 2007, **129**(5):1225-1232.
325. Carter DB, Chou KC: **A model for structure-dependent binding of Congo red to Alzheimer beta-amyloid fibrils.** *Neurobiol Aging* 1998, **19**(1):37-40.
326. Li L, Darden TA, Bartolotti L, Kominos D, Pedersen LG: **An atomic model for the pleated beta-sheet structure of Abeta amyloid protofilaments.** *Biophys J* 1999, **76**(6):2871-2878.
327. Klunk WE, Debnath ML, Pettegrew JW: **Development of small molecule probes for the beta-amyloid protein of Alzheimer's disease.** *Neurobiol Aging* 1994, **15**(6):691-698.
328. Cavillon F, Elhaddaoui A, Alix AJP, Turrell S, Dauchez M: **Identification of the importance of the secondary structure of Alzheimer's disease amyloid.** *J Mol Struct* 1997, **408**:185-189.
329. Groenning M, Olsen L, van de Weert M, Flink JM, Frokjaer S, Jorgensen FS: **Study on the binding of Thioflavin T to beta-sheet-rich and non-beta-sheet cavities.** *J Struct Biol* 2007, **158**(3):358-369.
330. Khurana R, Coleman C, Ionescu-Zanetti C, Carter SA, Krishna V, Grover RK, Roy R, Singh S: **Mechanism of thioflavin T binding to amyloid fibrils.** *J Struct Biol* 2005, **151**(3):229-238.

331. Skowronek M, Stopa B, Konieczny L, Rybarska J, Piekarska B, Szneler E, Bakalarski G, Roterman I: **Self-assembly of Congo Red - A theoretical and experimental approach to identify its supramolecular organization in water and salt solutions.** *Biopolymers* 1998, **46**(5):267-281.
332. Stopa B, Piekarska B, Konieczny L, Rybarska J, Spolnik P, Zemanek G, Roterman I, Krol M: **The structure and protein binding of amyloid-specific dye reagents.** *Acta Biochim Pol* 2003, **50**(4):1213-1227.
333. Roterman I, Krul M, Nowak M, Konieczny L, Rybarska J, Stopa B, Piekarska B, Zemanek G: **Why Congo red binding is specific for amyloid proteins - model studies and a computer analysis approach.** *Med Sci Monit* 2001, **7**(4):771-784.
334. Wu C, Wang ZX, Lei HX, Duan Y, Bowers MT, Shea JE: **The Binding of Thioflavin T and Its Neutral Analog BTA-1 to Protofibrils of the Alzheimer's Disease A beta(16-22) Peptide Probed by Molecular Dynamics Simulations.** *J Mol Biol* 2008, **384**(3):718-729.
335. Mishra R, Sellin D, Radovan D, Gohlke A, Winter R: **Inhibiting Islet Amyloid Polypeptide Fibril Formation by the Red Wine Compound Resveratrol.** *Chembiochem* 2009, **10**(3):445-+.
336. Cohen T, Frydman-Marom A, Rechter M, Gazit E: **Inhibition of amyloid fibril formation and cytotoxicity by hydroxyindole derivatives.** *Biochemistry* 2006, **45**(15):4727-4735.
337. Frid P, Anisimov SV, Popovic N: **Congo red and protein aggregation in neurodegenerative diseases.** *Brain Res Rev* 2007, **53**(1):135-160.
338. Chandler DE, Roberson RW: **Bioimaging : current concepts in light and electron microscopy.** Sudbury, Mass.: Jones and Bartlett Publishers; 2009.
339. Fandrich M, Meinhardt J, Grigorieff N: **Structural polymorphism of Alzheimer A beta and other amyloid fibrils.** *Prion* 2009, **3**(2):89-93.
340. Palmal S, Maity AR, Singh BK, Basu S, Jana NR, Jana NR: **Inhibition of amyloid fibril growth and dissolution of amyloid fibrils by curcumin-gold nanoparticles.** *Chemistry* 2014, **20**(20):6184-6191.

341. Giessibl FJ: **Theory for an Electrostatic Imaging Mechanism Allowing Atomic Resolution of Ionic-Crystals by Atomic Force Microscopy.** *Phys Rev B* 1992, **45**(23):13815-13818.
342. Binnig G, Quate CF, Gerber C: **Atomic Force Microscope.** *Phys Rev Lett* 1986, **56**(9):930-933.
343. Giessibl FJ: **Advances in atomic force microscopy.** *Rev Mod Phys* 2003, **75**(3):949-983.
344. Martin Y, Williams CC, Wickramasinghe HK: **Atomic Force Microscope Force Mapping and Profiling on a Sub 100-Å Scale.** *J Appl Phys* 1987, **61**(10):4723-4729.
345. Albrecht TR, Grutter P, Horne D, Rugar D: **Frequency-Modulation Detection Using High-Q Cantilevers for Enhanced Force Microscope Sensitivity.** *J Appl Phys* 1991, **69**(2):668-673.
346. Garcia R, Perez R: **Dynamic atomic force microscopy methods.** *Surf Sci Rep* 2002, **47**(6-8):197-301.
347. Anselmetti D, Luthi R, Meyer E, Richmond T, Dreier M, Frommer JE, Guntherodt HJ: **Attractive-Mode Imaging of Biological-Materials with Dynamic Force Microscopy.** *Nanotechnology* 1994, **5**(2):87-94.
348. Bustamante C, Keller D: **Scanning Force Microscopy in Biology.** *Phys Today* 1995, **48**(12):32-38.
349. San Paulo A, Garcia R: **High-resolution imaging of antibodies by tapping-mode atomic force microscopy: Attractive and repulsive tip-sample interaction regimes.** *Biophys J* 2000, **78**(3):1599-1605.
350. Reiter G, Castelein G, Hoerner P, Riess G, Sommer JU, Floudas G: **Morphologies of diblock copolymer thin films before and after crystallization.** *Eur Phys J E* 2000, **2**(4):319-334.
351. Zhong Q, Inniss D, Kjoller K, Elings VB: **Fractured Polymer Silica Fiber Surface Studied by Tapping Mode Atomic-Force Microscopy.** *Surf Sci* 1993, **290**(1-2):L688-L692.

352. Serpell LC, Sunde M, Benson MD, Tennent GA, Pepys MB, Fraser PE: **The protofilament substructure of amyloid fibrils.** *J Mol Biol* 2000, **300**(5):1033-1039.
353. Pedersen JS, Andersen CB, Otzen DE: **Amyloid structure--one but not the same: the many levels of fibrillar polymorphism.** *FEBS J* 2010, **277**(22):4591-4601.
354. Griffin MD, Mok ML, Wilson LM, Pham CL, Waddington LJ, Perugini MA, Howlett GJ: **Phospholipid interaction induces molecular-level polymorphism in apolipoprotein C-II amyloid fibrils via alternative assembly pathways.** *J Mol Biol* 2008, **375**(1):240-256.
355. Hatters DM, MacPhee CE, Lawrence LJ, Sawyer WH, Howlett GJ: **Human apolipoprotein C-II forms twisted amyloid ribbons and closed loops.** *Biochemistry* 2000, **39**(28):8276-8283.
356. Mangione P, Sunde M, Giorgetti S, Stoppini M, Esposito G, Gianelli L, Obici L, Asti L, Andreola A, Viglino P *et al*: **Amyloid fibrils derived from the apolipoprotein A1 Leu174Ser variant contain elements of ordered helical structure.** *Protein Sci* 2001, **10**(1):187-199.
357. Ahmad A, Millett IS, Doniach S, Uversky VN, Fink AL: **Partially folded intermediates in insulin fibrillation.** *Biochemistry* 2003, **42**(39):11404-11416.
358. Dong M, Hovgaard MB, Xu S, Otzen DE, Besenbacher F: **AFM study of glucagon fibrillation via oligomeric structures resulting in interwoven fibrils.** *Nanotechnology* 2006, **17**(16):4003-4009.
359. Greene LH, Grant TM: **Protein folding by 'levels of separation': A hypothesis.** *FEBS Lett* 2012, **586**(7):962-966.
360. Higman VA, Greene LH: **Elucidation of conserved long-range interaction networks in proteins and their significance in determining protein topology.** *Physica A* 2006, **368**(2):595-606.
361. Kinch LN, Grishin NV: **Evolution of protein structures and functions.** *Curr Opin Struct Biol* 2002, **12**(3):400-408.

362. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic Local Alignment Search Tool.** *J Mol Biol* 1990, **215**(3):403-410.
363. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**(17):3389-3402.
364. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF: **Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements.** *Nucleic Acids Res* 2001, **29**(14):2994-3005.
365. Todd AE, Orengo CA, Thornton JM: **Evolution of function in protein superfamilies, from a structural perspective.** *J Mol Biol* 2001, **307**(4):1113-1143.
366. Alexander RP, Zhulin IB: **Evolutionary genomics reveals conserved structural determinants of signaling and adaptation in microbial chemoreceptors.** *Proc Natl Acad Sci U S A* 2007, **104**(8):2885-2890.
367. Murzin AG: **Biochemistry - Metamorphic proteins.** *Science* 2008, **320**(5884):1725-1726.
368. Weissmann C: **Birth of a prion: Spontaneous generation revisited.** *Cell* 2005, **122**(2):165-168.
369. Tuinstra RL, Peterson FC, Kutlesa S, Elgin ES, Kron MA, Volkman BF: **Interconversion between two unrelated protein folds in the lymphotactin native state.** *Proc Natl Acad Sci* 2008, **105**(13):5057-5062.
370. Luo XL, Yu HT: **Protein Metamorphosis: The Two-State Behavior of Mad2.** *Structure* 2008, **16**(11):1616-1625.
371. Mapelli M, Musacchio A: **MAD contortions: conformational dimerization boosts spindle checkpoint signaling.** *Curr Opin Struct Biol* 2007, **17**(6):716-725.
372. Littler DR, Harrop SJ, Fairlie WD, Brown LJ, Pankhurst GJ, Pankhurst S, DeMaere MZ, Campbell TJ, Bauskin AR, Tonini R *et al*: **The intracellular**

chloride ion channel protein CLIC1 undergoes a redox-controlled structural transition. *J Biol Chem* 2004, **279**(10):9298-9305.

373. He YA, Chen YH, Alexander PA, Bryan PN, Orban J: **Mutational Tipping Points for Switching Protein Folds and Functions.** *Structure* 2012, **20**(2):283-291.
374. Roessler CG, Hall BM, Anderson WJ, Ingram WM, Roberts SA, Montfort WR, Cordes MHJ: **Transitive homology-guided structural studies lead to discovery of Cro proteins with 40% sequence identity but different folds.** *Proc Natl Acad Sci* 2008, **105**(7):2343-2348.
375. Belogurov GA, Mooney RA, Svetlov V, Landick R, Artsimovitch I: **Functional specialization of transcription elongation factors.** *Embo J* 2009, **28**(2):112-122.
376. Bryan PN, Orban J: **Proteins that switch folds.** *Curr Opin Struct Biol* 2010, **20**(4):482-488.
377. Ambroggio XI, Kuhlman B: **Design of protein conformational switches.** *Curr Opin Struct Biol* 2006, **16**(4):525-530.
378. Ambroggio XI, Kuhlman B: **Computational design of a single amino acid sequence that can switch between two distinct protein folds.** *J Am Chem Soc* 2006, **128**(4):1154-1161.
379. Dalal S, Regan L: **Understanding the sequence determinants of conformational switching using protein design.** *Protein Sci* 2000, **9**(9):1651-1659.
380. Blanco FJ, Angrand I, Serrano L: **Exploring the conformational properties of the sequence space between two proteins with different folds: An experimental study.** *J Mol Biol* 1999, **285**(2):741-753.
381. Rose GD, Creamer TP: **Protein-Folding - Predicting Predicting.** *Proteins-Structure Function and Genetics* 1994, **19**(1):1-3.
382. He YN, Rozak DA, Sari N, Chen YH, Bryan P, Orban J: **Structure, dynamics, and stability variation in bacterial albumin binding modules: Implications for species specificity.** *Biochemistry* 2006, **45**(33):10102-10109.

383. Fahnestock SR, Alexander P, Nagle J, Filpula D: **Gene for an Immunoglobulin-Binding Protein from a Group-G Streptococcus.** *J Bacteriol* 1986, **167**(3):870-880.
384. Johansson MU, deChateau M, Wikstrom M, Forsen S, Drakenberg T, Bjorck L: **Solution structure of the albumin-binding GA module: A versatile bacterial protein domain.** *J Mol Biol* 1997, **266**(5):859-865.
385. Falkenberg C, Bjorck L, Akerstrom B: **Localization of the Binding-Site for Streptococcal Protein-G on Human Serum-Albumin - Identification of a 5.5-Kilodalton Protein-G Binding Albumin Fragment.** *Biochemistry* 1992, **31**(5):1451-1457.
386. Gronenborn AM, Filpula DR, Essig NZ, Achari A, Whitlow M, Wingfield PT, Clore GM: **A Novel, Highly Stable Fold of the Immunoglobulin Binding Domain of Streptococcal Protein-G.** *Science* 1991, **253**(5020):657-661.
387. Myhre EB, Kronvall G: **Heterogeneity of Nonimmune Immunoglobulin-Fc Reactivity among Gram-Positive Cocci - Description of 3 Major Types of Receptors for Human Immunoglobulin-G.** *Infect Immun* 1977, **17**(3):475-482.
388. Alexander PA, He Y, Chen Y, Orban J, Bryan PN: **The design and characterization of two proteins with 88% sequence identity but different structure and function.** *Proc Natl Acad Sci* 2007, **104**(29):11963-11968.
389. Alexander PA, He Y, Chen Y, Orban J, Bryan PN: **A minimal sequence code for switching protein structure and function.** *Proc Natl Acad Sci* 2009, **106**(50):21149-21154.
390. He Y, Chen YH, Alexander P, Bryan PN, Orban J: **NMR structures of two designed proteins with high sequence identity but different fold and function.** *Proc Natl Acad Sci* 2008, **105**(38):14412-14417.
391. Bae T, Schneewind O: **The YSIRK-G/S motif of staphylococcal protein A and its role in efficiency of signal peptide processing.** *J Bacteriol* 2003, **185**(9):2910-2919.
392. Finn RD, Bateman A, Clements J, Coghill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J *et al*: **Pfam: the protein families database.** *Nucleic Acids Res* 2014, **42**(Database issue):D222-230.

- 393. Sonnhammer EL, Eddy SR, Durbin R: **Pfam: a comprehensive database of protein domain families based on seed alignments.** *Proteins* 1997, **28**(3):405-420.
- 394. Schmohl L, Schwarzer D: **Sortase-mediated ligations for the site-specific modification of proteins.** *Curr Opin Chem Biol* 2014, **22C**:122-128.
- 395. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L *et al*: **SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information.** *Nucleic Acids Res* 2014, **42**(W1):W252-W258.
- 396. Skolnick J, Kihara D, Zhang Y: **Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm.** *Proteins-Structure Function and Bioinformatics* 2004, **56**(3):502-518.
- 397. MacKenzie DA, Tailford LE, Hemmings AM, Juge N: **Crystal Structure of a Mucus-binding Protein Repeat Reveals an Unexpected Functional Immunoglobulin Binding Activity.** *J Biol Chem* 2009, **284**(47):32444-32453.
- 398. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32**(5):1792-1797.
- 399. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: **MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0.** *Mol Biol Evol* 2013, **30**(12):2725-2729.
- 400. Nielsen M, Lundegaard C, Lund O, Petersen TN: **CPHmodels-3.0-remote homology modeling using structure-guided sequence profiles.** *Nucleic Acids Res* 2010, **38**:W576-W581.
- 401. Schueler-Furman O, Baker D: **Conserved residue clustering and protein structure prediction.** *Proteins* 2003, **52**(2):225-235.
- 402. Greene LH, Hamada D, Eyles SJ, Brew K: **Conserved signature proposed for folding in the lipocalin superfamily.** *FEBS Lett* 2003, **553**(1-2):39-44.
- 403. Greene LH, Higman VA: **Uncovering network systems within protein structures.** *J Mol Biol* 2003, **334**(4):781-791.

- 404. Greene LH, Higman VA: **Conserved networks and the determinants of protein topology.** *FEBS Journal* 2005, **272**:87-87.
- 405. Li H, Wojtaszek JL, Greene LH: **Analysis of conservation in the Fas-associated death domain protein and the importance of conserved tryptophans in structure, stability and folding.** *BBA-Proteins Proteom* 2009, **1794**(4):583-593.
- 406. Geierhaas CD, Paci E, Vendruscolo M, Clarke J: **Comparison of the transition states for folding of two Ig-like proteins from different superfamilies.** *J Mol Biol* 2004, **343**(4):1111-1123.
- 407. Martinez JC, Serrano L: **The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved.** *Nat Struct Biol* 1999, **6**(11):1010-1016.
- 408. Mirny LA, Shakhnovich EI: **Universally conserved positions in protein folds: Reading evolutionary signals about stability, folding kinetics and function.** *J Mol Biol* 1999, **291**(1):177-196.
- 409. Kloczkowski A, Jernigan RL: **Loop folds in proteins and evolutionary conservation of folding nuclei.** *J Biomol Struct Dyn* 2002, **20**(3):323-325.
- 410. Ting KLH, Jernigan RL: **Identifying a folding nucleus for the lysozyme/alpha-lactalbumin family from sequence conservation clusters.** *J Mol Evol* 2002, **54**(4):425-436.
- 411. Gunasekaran K, Hagler AT, Gierasch LM: **Sequence and structural analysis of cellular retinoic acid-binding proteins reveals a network of conserved hydrophobic interactions.** *Proteins-Structure Function and Genetics* 2004, **54**(2):179-194.
- 412. Guo Y, Yu XM, Rihani K, Wang QY, Rong LJ: **The role of a conserved acidic residue in calcium-dependent protein folding for a low density lipoprotein (LDL)-A module - Implications in structure and function for the LDL receptor superfamily.** *J Biol Chem* 2004, **279**(16):16629-16637.
- 413. Guo ZY, Wang S, Tang YH, Feng YM: **Mutagenesis of the three conserved valine residues: consequence on the foldability of insulin.** *BBA - Proteins Proteom* 2004, **1699**(1-2):103-109.

- 414. Pearce MC, Cabrita LD, Ellisdon AM, Bottomley SP: **The loss of tryptophan 194 in antichymotrypsin lowers the kinetic barrier to misfolding.** *Febs Journal* 2007, **274**(14):3622-3632.
- 415. Kragelund BB, Poulsen K, Andersen KV, Baldursson T, Kroll JB, Neergard TB, Jepsen J, Roepstorff P, Kristiansen K, Poulsen FM *et al*: **Conserved residues and their role in the structure, function, and stability of acyl-coenzyme A binding protein.** *Biochemistry* 1999, **38**(8):2386-2394.
- 416. Hamill SJ, Steward A, Clarke J: **The folding of an immunoglobulin-like Greek key protein is defined by a common-core nucleus and regions constrained by topology.** *J Mol Biol* 2000, **297**(1):165-178.
- 417. Fowler SB, Clarke J: **Mapping the folding pathway of an immunoglobulin domain: Structural detail from phi value analysis and movement of the transition state.** *Structure* 2001, **9**(5):355-366.
- 418. Heidary DK, Jennings PA: **Three topologically equivalent core residues affect the transition state ensemble in a protein folding reaction.** *J Mol Biol* 2002, **316**(3):789-798.
- 419. Otzen DE, Oliveberg M: **Conformational plasticity in folding of the split beta-alpha-beta protein S6: Evidence for burst-phase disruption of the native state.** *J Mol Biol* 2002, **317**(4):613-627.
- 420. Hubner IA, Oliveberg M, Shakhnovich EI: **Simulation, experiment, and evolution: Understanding nucleation in protein S6 folding.** *Proc Natl Acad Sci* 2004, **101**(22):8354-8359.
- 421. Wilson CJ, Wittung-Stafshede P: **Role of structural determinants in folding of the sandwich-like protein *Pseudomonas aeruginosa* azurin.** *Proc Natl Acad Sci* 2005, **102**(11):3984-3987.
- 422. Campbell-Valois FX, Michnick SW: **The transition state of the ras binding domain of raf is structurally polarized based on phi-values but is energetically diffuse.** *J Mol Biol* 2007, **365**(5):1559-1577.
- 423. Olofsson M, Hansson S, Hedberg L, Logan DT, Oliveberg M: **Folding of S6 structures with divergent amino acid composition: Pathway flexibility within partly overlapping foldons.** *J Mol Biol* 2007, **365**(1):237-248.

424. Lappalainen I, Hurley MG, Clarke J: **Plasticity within the obligatory folding nucleus of an immunoglobulin-like domain.** *J Mol Biol* 2008, **375**(2):547-559.
425. Larson SM, Ruczinski I, Davidson AR, Baker D, Plaxco KW: **Residues participating in the protein folding nucleus do not exhibit preferential evolutionary conservation.** *J Mol Biol* 2002, **316**(2):225-233.
426. Tseng YY, Liang J: **Are residues in a protein folding nucleus evolutionarily conserved?** *J Mol Biol* 2004, **335**(4):869-880.
427. Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R: **Evolutionary information for specifying a protein fold.** *Nature* 2005, **437**(7058):512-518.
428. Marcelino AMC, Smock RG, Gierasch LM: **Evolutionary coupling of structural and functional sequence information in the intracellular lipid-binding protein family.** *Proteins-Structure Function and Bioinformatics* 2006, **63**(2):373-384.
429. Mullis KB: **The Unusual Origin of the Polymerase Chain-Reaction.** *Sci Am* 1990, **262**(4):56-&.
430. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA: **Primer-Directed Enzymatic Amplification of DNA with a Thermostable DNA-Polymerase.** *Science* 1988, **239**(4839):487-491.
431. Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N: **Enzymatic Amplification of Beta-Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle-Cell Anemia.** *Science* 1985, **230**(4732):1350-1354.
432. Erijman A, Dantes A, Bernheim R, Shifman JM, Peleg Y: **Transfer-PCR (TPCR): A highway for DNA cloning and protein engineering.** *J Struct Biol* 2011, **175**(2):171-177.
433. Erlich HA, Gelfand D, Sninsky JJ: **Recent Advances in the Polymerase Chain-Reaction.** *Science* 1991, **252**(5013):1643-1651.

- 434. Holm L, Rosenstrom P: **Dali server: conservation mapping in 3D.** *Nucleic Acids Res* 2010, **38**:W545-W549.
- 435. Holm L, Kaariainen S, Rosenstrom P, Schenkel A: **Searching protein structure databases with DaliLite v.3.** *Bioinformatics* 2008, **24**(23):2780-2781.
- 436. Holm L, Park J: **DaliLite workbench for protein structure comparison.** *Bioinformatics* 2000, **16**(6):566-567.
- 437. Rost B: **Twilight zone of protein sequence alignments.** *Protein Eng* 1999, **12**(2):85-94.
- 438. O'Mara D, Vershon AK, Herman T, Conti B, Fernando M, Lashley D, Li A, Paton S, White J, Horlbeck M: **Characterization of sirtuin proteins using RasMol-RP and three dimensional modeling.** *FASEB Journal* 2007, **21**(6):A1036-A1036.
- 439. Pembroke JT: **Bio-molecular modelling utilising RasMol and PDB resources: a tutorial with HEW lysozyme.** *Biochem Mol Biol Edu* 2000, **28**(6):297-300.
- 440. Horton RM, Stone RJS: **An introduction to molecular visualization: Seeing in stereo with RasMol.** *Biotechniques* 1997, **22**(4):660-&.
- 441. Sayle RA, Milnerwhite EJ: **Rasmol - Biomolecular Graphics for All.** *Trends Biochem Sci* 1995, **20**(9):374-376.
- 442. Gallivan JP, Dougherty DA: **Cation-pi interactions in structural biology.** *Proc Natl Acad Sci* 1999, **96**(17):9459-9464.
- 443. Chen CC, Hsu W, Hwang KC, Hwu JR, Lin CC, Horng JC: **Contributions of cation-pi interactions to the collagen triple helix stability.** *Arch Biochem Biophys* 2011, **508**(1):46-53.
- 444. Gallivan JP, Dougherty DA: **Cation-pi interactions as structural motifs in proteins.** *Abstr Pap Am Chem S* 1999, **217**:U87-U87.
- 445. Thurman R, Rajalingam D, Kumar TKS: **Cation-Pi Interactions Contribute Significantly to the Stability of FGF and the FGFR.** *Biophys J* 2010, **98**(3):447a-448a.

- 446. Wintjens R, Lievin J, Rooman M, Buisine E: **Contribution of cation-pi interactions to the stability of protein-DNA complexes.** *J Mol Biol* 2000, **302**(2):395-410.
- 447. Gasyimov OK, Abduragimov AR, Glasgow BJ: **Cation-pi Interactions in Lipocalins: Structural and Functional Implications.** *Biochemistry* 2012, **51**(14):2991-3002.
- 448. Hasegawa H, Holm L: **Advances and pitfalls of protein structural alignment.** *Curr Opin Struct Biol* 2009, **19**(3):341-348.
- 449. Sander C, Schneider R: **Database of Homology-Derived Protein Structures and the Structural Meaning of Sequence Alignment.** *Proteins-Structure Function and Genetics* 1991, **9**(1):56-68.
- 450. Kyte J, Doolittle RF: **A Simple Method for Displaying the Hydropathic Character of a Protein.** *J Mol Biol* 1982, **157**(1):105-132.
- 451. Gill SC, Vonhippel PH: **Calculation of Protein Extinction Coefficients from Amino-Acid Sequence Data.** *Anal Biochem* 1989, **182**(2):319-326.
- 452. Shi XS, Bisaria N, Benz-Moy TL, Bonilla S, Pavlichin DS, Herschlag D: **Roles of Long-Range Tertiary Interactions in Limiting Dynamics of the Tetrahymena Group I Ribozyme.** *J Am Chem Soc* 2014, **136**(18):6643-6648.
- 453. Tanaka S, Scheraga HA: **Model of protein folding: inclusion of short-, medium-, and long-range interactions.** *Proc Natl Acad Sci U S A* 1975, **72**(10):3802-3806.
- 454. Havlin RH, Tycko R: **Probing site-specific conformational distributions in protein folding with solid-state NMR.** *Proc Natl Acad Sci U S A* 2005, **102**(9):3284-3289.
- 455. Hu KN, Yau WM, Tycko R: **Detection of a transient intermediate in a rapid protein folding process by solid-state nuclear magnetic resonance.** *J Am Chem Soc* 2010, **132**(1):24-25.
- 456. Thurber KR, Tycko R: **Biomolecular solid state NMR with magic-angle spinning at 25K.** *J Magn Reson* 2008, **195**(2):179-186.

- 457. Balbach J, Forge V, Lau WS, vanNuland NAJ, Brew K, Dobson CM: **Protein folding monitored at individual residues during a two-dimensional NMR experiment.** *Science* 1996, **274**(5290):1161-1163.
- 458. Eliezer D, Yao J, Dyson HJ, Wright PE: **Structural and dynamic characterization of partially folded states of apomyoglobin and implications for protein folding.** *Nat Struct Biol* 1998, **5**(2):148-155.
- 459. Freund SMV, Wong KB, Fersht AR: **Initiation sites of protein folding by NMR analysis.** *Proc Natl Acad Sci* 1996, **93**(20):10600-10603.
- 460. Dyson HJ, Wright PE: **Unfolded proteins and protein folding studied by NMR.** *Chem Rev* 2004, **104**(8):3607-3622.
- 461. Rosen LE, Connell KB, Marqusee S: **Evidence for close side-chain packing in an early protein folding intermediate previously assumed to be a molten globule.** *Proc Natl Acad Sci* 2014, **111**(41):14746-14751.
- 462. Park SH, O'Neil KT, Roder H: **An early intermediate in the folding reaction of the B1 domain of protein G contains a native-like core.** *Biochemistry* 1997, **36**(47):14277-14283.
- 463. Barton D, Nakanishi KJ, Meth-Cohn O: **Comprehensive natural products chemistry**, 1st edn. Amsterdam ; New York: Elsevier; 1999.
- 464. Bell CE, Lewis M: **A closer view of the conformation of the Lac repressor bound to operator.** *Nat Struct Biol* 2000, **7**(3):209-214.
- 465. Daber R, Stayrook S, Rosenberg A, Lewis M: **Structural analysis of lac repressor bound to allosteric effectors.** *J Mol Biol* 2007, **370**(4):609-619.
- 466. Primrose SB, Twyman R: **Principles of gene manipulation and genomics**: John Wiley & Sons; 2013.
- 467. Saitô H, Ando I, Naito A: **Solid state NMR spectroscopy for biopolymers : principles and applications.** Dordrecht: Springer; 2006.

468. Uversky VN: **Neuropathology, biochemistry, and biophysics of alpha-synuclein aggregation.** *J Neurochem* 2007, **103**(1):17-37.
469. Straub JE, Thirumalai D: **Toward a molecular theory of early and late events in monomer to amyloid fibril formation.** *Annu Rev Phys Chem* 2011, **62**:437-463.
470. Ahmad B, Winkelmann J, Tiribilli B, Chiti F: **Searching for conditions to form stable protein oligomers with amyloid-like characteristics: The unexplored basic pH.** *BBA* 2010, **1804**(1):223-234.
471. Sicorello A, Torrassa S, Soldi G, Gianni S, Travaglini-Allocatelli C, Taddei N, Relini A, Chiti F: **Agitation and high ionic strength induce amyloidogenesis of a folded PDZ domain in native conditions.** *Biophys J* 2009, **96**(6):2289-2298.
472. Wu WH, Sun X, Yu YP, Hu J, Zhao L, Liu Q, Zhao YF, Li YM: **TiO₂ nanoparticles promote beta-amyloid fibrillation in vitro.** *Biochem Biophys Res Commun* 2008, **373**(2):315-318.
473. Fezoui Y, Teplow DB: **Kinetic studies of amyloid beta-protein fibril assembly. Differential effects of alpha-helix stabilization.** *J Biol Chem* 2002, **277**(40):36948-36954.
474. Motamedi-Shad N, Monsellier E, Torrassa S, Relini A, Chiti F: **Kinetic analysis of amyloid formation in the presence of heparan sulfate: faster unfolding and change of pathway.** *J Biol Chem* 2009, **284**(43):29921-29934.
475. Guijarro JI, Sunde M, Jones JA, Campbell ID, Dobson CM: **Amyloid fibril formation by an SH3 domain.** *Proc Natl Acad Sci* 1998, **95**(8):4224-4228.
476. Krebs MR, Devlin GL, Donald AM: **Amyloid fibril-like structure underlies the aggregate structure across the pH range for beta-lactoglobulin.** *Biophys J* 2009, **96**(12):5013-5019.
477. Chiti F, Bucciantini M, Capanni C, Taddei N, Dobson CM, Stefani M: **Solution conditions can promote formation of either amyloid protofilaments or mature fibrils from the HypF N-terminal domain.** *Protein Sci* 2001, **10**(12):2541-2547.

478. Berglund H, Olerenshaw D, Sankar A, Federwisch M, McDonald NQ, Driscoll PC: **The three-dimensional solution structure and dynamic properties of the human FADD death domain.** *J Mol Biol* 2000, **302**(1):171-188.
479. Grothe HL, Little MR, Cho AS, Huang AJ, Yuan C: **Denaturation and solvent effect on the conformation and fibril formation of TGFBIp.** *Mol Vis* 2009, **15**:2617-2626.
480. Nilsson MR: **Techniques to study amyloid fibril formation in vitro.** *Methods* 2004, **34**(1):151-160.
481. Eisert R, Felau L, Brown LR: **Methods for enhancing the accuracy and reproducibility of Congo red and thioflavin T assays.** *Anal Biochem* 2006, **353**(1):144-146.
482. Fandrich M, Zandomenighi G, Krebs MR, Kittler M, Buder K, Rossner A, Heinemann SH, Dobson CM, Diekmann S: **Apomyoglobin reveals a random-nucleation mechanism in amyloid protofibril formation.** *Acta Histochem* 2006, **108**(3):215-219.
483. Fei L, Perrett S: **Disulfide bond formation significantly accelerates the assembly of Ure2p fibrils because of the proximity of a potential amyloid stretch.** *J Biol Chem* 2009, **284**(17):11134-11141.

APPENDIX I

BIOINFORMATICS OF GA AND GB1 EVOLUTION

WT-GB1-YSIRK	-----MVGKNNYKTRTEKAAN-----KKQRFSLKKLSVGVASVAVGTTLF-LS
WT-GA-YSIRK	-----MEN-----KKMKYYLRKSAFGLAAV--SASVL-VG
YP_005861639.1	-----MGMFFNKKDND-----SKQRFGRILTIAACSLLSTLILGIG
WP_021874218.1	-----MVSKNNRDKKMEAVAE-----RKPFAIRILTIAAASLLGTSLW-MS
KFL97016.1	MIYYDSKWRETGMLSKNNYQERLRKMDD-----KQERFSIRFSV AAS LVGTAIL-SM
KFL95230.1	MIYYDSKWRETGMLSKNNYQERLRKMDD-----KQERFSIRFSV AAS LVGTAIL-SM
WP_033688740.1	-----MYKSKRRKNKSFWDYGLSQRFIRYHF AAS LLGTALI-LG
WP_003097688.1	-----MEKFHGR-----KAQRFIRYSF AAS LLGTALF-LG
WP_033689127.1	-----MYSRMEKYHGR-----RAQRFIRYSF AAS LLGTALL-LG
WP_020903468.1	-----MYSRMEKYHGR-----RAQRFIRYSF AAS LLGTALV-LG
WP_000287308.1	-----MYSRMEKYHGR-----RAQRFIRYSF AAS LLGTALL-LG
WP_033685380.1	-----MYSRMEKYHGR-----RAQRFIRYSF AAS LLGTALV-LG
WP_033681547.1	-----MYSRMEKYHGR-----RAQRFIRYSF AAS LLGTALL-LG
WP_023948409.1	-----MYSRMDKYHGQ-----KVQRFIRYSF AAS LLGTALL-LG
WT-GB1-YSIRK	NT-----DAVSA-----E
WT-GA-YSIRK	VT--TVSAQVTTR-----A
YP_005861639.1	TQEONSKAQAAATTETSNTASSTDLVDNHRNKNTY-----L
WP_021874218.1	TSTSTVHADETDNNDSDAKTNLESNQASSTGHVEKVVVEQNQTANENTDDSTKTNNVSAQ
KFL97016.1	QNVQTVHADATTDTEKGTDTVTSKNDEQNKQKAYNQVVSSED-----Q
KFL95230.1	QNVQTVRADATTDTEKGTDTVTSKNDEQNKQKAYNQVVSSED-----Q
WP_033688740.1	AA--QTAKAEETVTENKTEAVASAPKDDKASENVNTVTPALSATTEAAVVEKPTLSDE
WP_003097688.1	AN--GVRADEATPSVNPATSGLSNSDKNVSGS---TLSTPVV-----E
WP_033689127.1	AN--AVKADETLV-VNPTASDLAATNKKDADS---ALTTPVV-----E
WP_020903468.1	AN--GVQAEETVA-VNPATSELSNSDKNLSGS---TLSTPVV-----E
WP_000287308.1	AN--AVKADETS-ASTKTSEVTNSDKQKPDSD---AITTPVV-----E
WP_033685380.1	AN--GVQADETLV-VNPATSDLAATNKKDADS---ALTTPVV-----E
WP_033681547.1	AN--AVKADETS-ASAKTPEVTNSDKQKSDSD---GSTTPVV-----E
WP_023948409.1	TN--AVKADETNT-GSAKATEITNPDQKQKPDSD---AITTPVV-----E
WT-GB1-YSIRK	ENPDRNIEQLQAAL-----TEKE
WT-GA-YSIRK	QAAKLREEATQKISE-----
YP_005861639.1	SSSEVNETATKVNEKETSAAGNEQQDSQSAVAQDKQSGEKAADVVTNNRSTVEDKSSNNAE
WP_021874218.1	NTQESVDESSDISSD-NAQQNKAITSEEQNSDAAVTIDNNQAADENKAETQKVTDKTTKT
KFL97016.1	NKASKTTDTTMVGQDSKVASFSASKNEGTFAEASSEDKTSSTTDATQNKASENTKATEDN
KFL95230.1	NKASKTTDTTMVGQDSKVASFSASKNEGTFAEASSEDKTSSTTDATQNKASENTKATEDN
WP_033688740.1	EVAKLAAEASKKDDK-AS--ETATTEKTEAADKEKATLT--APLTDKK-----ADKAVDE
WP_003097688.1	KLPELKIDAVKADEN-AEVKEESKNEVTTVAEKEVTEAT--TDKTDKKTETDVKEKSDKE
WP_033689127.1	ELPELKIDAVKADEK-AKAKEDAKTEATPVVEKEITEAT--AEKTDKKLETDVKEKSDKE
WP_020903468.1	ELPELKIDVKAEEK-TEAKEDAKTAATPVAEKEVTEAT--ADKTDKKLETDVKEKSDKE
WP_000287308.1	ELPELKIDAVKADEK-PEVKEEAK-----PVAEKEVTDKA--A-----TEKSDKE
WP_033685380.1	ELPELKIDAVKADEK-AEPKEDVKTEATPVVEKEVSDKS--DKEAS-----KEKSDKE
WP_033681547.1	ELPELKIDAVKADEK-PEVKEEAKTEAKPVAEKEVADK-----AA-----TEKSDKE
WP_023948409.1	ELPELKIDAVKADEK-SEVKE-----EAKPVAEKEVTDK-----AA-----TEKSDKE

Figure A1. MUSCLE alignment of 12 mucus-binding proteins with YSIRK family proteins. Positions in red indicate positions of complete conservation and position in blue are positions conserved in amino acid character. Accession numbers correspond to mucus-binding proteins found in Figure 40.

WT-GB1-YSIRK	FAIAEVQAAGFTDQKYVDWINA-QETVQDVN-----GVKREILTTVPDEEQS-
WT-GA-YSIRK	-LEKTIDNKTQLD-----SVKKEIQNAVDRDKIQ-
YP_005861639.1	SSETTDNTKNTVNSDRAEVTNKEANTQTDSKKVTQKTSQAVNDVNKNVAETTTDTRNTK
WP_021874218.1	KQDDNKSSQTIDNKKSSSEKAATDTSNKNNVEQSANSVENNA--NIDNSIAANTQTDITKS
KFL97016.1	KVDAVADKSTEDKNTATTQESSDNKSTENKTTDAQKVESKVATAKATTTDANSVKKTAS-
KFL95230.1	KVDAVADKSTEDKNTATTQESSDNKSTENKTTDAQKVESKVATAKATTTDANSVKKTAS-
WP_033688740.1	KADKKDEKK-----A-ENPITATKTVLEQLTSEA--EVLNTTASNFAADKKAE-
WP_003097688.1	HADKTEADKEKT-----EKVET-EKAQDDVKT VLTQLTSEA--DVMATVASNFSDKEVT-
WP_033689127.1	QADKKEATKEKTDKKTSEKVT-EKVQDDVKT VLTQLTSEA--DVMASVASNFSDKEVK-
WP_020903468.1	QSDKKEADKEKTNKETSEKVT-EKAQDDVKT VLTQLTSEA--DVMATVASNFSDKEVK-
WP_000287308.1	QADKKEVAKEKTDKESPKKAAT-EKAQDEVKT VLTQLTSEA--EVMASVASNFSDKEVK-
WP_033685380.1	QADKKEATKEKTDKETSEKVT-EKAQDDVKT VLTQLTSEA--EVMATVASNFSDKEAK-
WP_033681547.1	QADKKEADKEKTNKETSEKATT-EKAQDEVKT VLTQLTSEA--EVMASVASNFSDKEVK-
WP_023948409.1	QADKKEVAKEKTDKETSEKAET-VKPKDEVKT VLTQLTSEA--EVMATVASNFSDKEAK-
WT-GB1-YSIRK	-----GIEEETSVDSESDFNAPDFDW
WT-GA-YSIRK	-----ELSNQA-----
YP_005861639.1	VSSYTSNLDL-----ENIQESLEQQAKENNGKALDA
WP_021874218.1	NIQLNESLPSIAQAGQNGKTIVKDNDDTTQELKIGDLSSDLSGDALKANLTGKNQVLLNQ
KFL97016.1	-----TTSTDQ--TTSVNTTTTFTNTQ
KFL95230.1	-----TTSTDQ--TTSVNTTTTFTNTQ
WP_033688740.1	-----DKAGKE--AIATAVASAKVQI
WP_003097688.1	-----GVEDKQ--NLSAAITAVKLEA
WP_033689127.1	-----DDASKQ--KLSAAVAAVKLEA
WP_020903468.1	-----DVESKQ--KLSAAIAAVKLEA
WP_000287308.1	-----DEAAKK--ELAVTIEAVKLEA
WP_033685380.1	-----DVEAKQ--KLSAAIAAVKLEA
WP_033681547.1	-----DEAAKK--ELAVKIEAVKLEA
WP_023948409.1	-----DVEAKQ--KLSAAIAAVKLEA
WT-GB1-YSIRK	-----SGNDEAVAAEA-----ELQAAKESAIAEVKAAGFTDQ-----
WT-GA-YSIRK	-----DQIV-----SAQAEKEAL-----
YP_005861639.1	KSVTSLLRSSD---AANFQVLAALTENIAEADKIKANAAVTIPSTDGRYTLYISRNTWGN
WP_021874218.1	SNSSEVVVGKNVDPTKQLQAMARTA-----MFAAVNPNAADNYTTVSDFNALQQAV
KFL97016.1	SSTAVLFSASA---LSESKALAATP-----RASQATTNAQAKNNNYKLVTSSSELQ
KFL95230.1	SSTAALFSASA---LSESKALAATP-----RASSATTNAQAKNNNYKLVTSSSELQ
WP_033688740.1	EASKKALAAGE---ITKQELDAQLQ-----RISSAIEAVYDEMKRAGHLGKVEAVL
WP_003097688.1	TASKELL-LSD---ASKDQMVAQVN-----RLSDAIEAVYAEMKRAGHAGKVEAVI
WP_033689127.1	VASKGLL-SSD---ASKDQMVAQVN-----RLSAAIEAVYAEMKRAGHAGKVESVL
WP_020903468.1	VASKGLL-SSD---ASKDQMVAQVN-----RLSAAIEAVYAEMKRAGHAGKVEAVL
WP_000287308.1	AKSNDLL-SSD---ASKDQMVAQVN-----RLSAAIEAVYTEMKRAGHAGKVESVL
WP_033685380.1	VAAKGLLYSND---STEEQLTSQVN-----RISSAIEAVYAEMKRAGHAGKVEAVL
WP_033681547.1	AKSNDLL-SSD---ASKDQMVAQVN-----RLSAAIEAVYAEMKRAGHAGKVEARL
WP_023948409.1	VASKGLL-SSD---ASKDQMVAQVN-----RLSAAIEAVYAEMKRAGHAGKVEAVL

Figure A1. Continued.

WT-GB1-YSIRK	-----KYVDW-----	INAQE-TVQDVNGV-----
WT-GA-YSIRK	-----	IKSEK-KLADAWEL-----
YP_005861639.1	TTDGNQPTKVLLSGNVLSGDTVVISIPSYGI-----	VGVNSPTIDANYGSASLKDMGN-
WP_021874218.1	NDYSVSGVNIISGNITAYGD---LNINRTFTI-----	KGADN-NATLSLGQNKINNNQGL
KFL97016.1	QAINSGVAGINIDRSIDASNVNLAITNTFAI-----	VGIND-AAVLNLGQKSLNNSGNL
KFL95230.1	QAINSGVAGINIDRSIDASNVNLAITNTFAI-----	VGIND-AAVLNLGQKSLNNSGNL
WP_033688740.1	AGETTTNTAIVAPTPTKTPVKNINKLTDEEIAAVKREIMNANP-SITDPSMI-----	
WP_003097688.1	DDATTQDQKIVGKDVVKDGQKVSVTNGYIT-----	MNADN-TAPKAWGF-----
WP_033689127.1	ADTAS---KITGKDILRDGETVNAVNTAYVD-----	MNADN-TAPTGWGF-----
WP_020903468.1	ADTAS---KITGKDVLDGETVNAVNTAYVE-----	MNADN-TRPTGWGF-----
WP_000287308.1	AATAS---KITGKDVLDGETVNAVNTAYVD-----	MNADN-TKPVGWGF-----
WP_033685380.1	GATDAKDQTITGKGVIRNGNAIVTVNNAYVT-----	MNADN-TAPKAWGI-----
WP_033681547.1	AETH-----TGKALIKEGKAVNVQNAYIT-----	MNADN-SAPTAWGI-----
WP_023948409.1	AETH-----TGKALIKEGKAVNVQNAYIT-----	MNADN-SAPTAWGI-----
WT-GB1-YSIRK	---KREILTT-----	VPDEEQSGI-----
WT-GA-YSIRK	---EAAKDAA-----	LKELNQYGV-----
YP_005861639.1	---DKVVIYNFTTSGV-----	INPIIT-----IPADNGYGA-----
WP_021874218.1	TLDDITVNGSILGNGTVNIKGTVTNNVSVNSS-----	VPTQDQFKAQNYTGN
KFL97016.1	TLQDITINGAVSGNGTVNIKGNVTSNVNENNSLIKGATADAANAALKDQTSTGTIGTQGT	
KFL95230.1	TLQDITINGAVSGNGTVNIKGNVTSNVNENNSLIKGATADAANAALKDQTSTGTIGTQGT	
WP_033688740.1	---EVVQDGNGTAGGA-----	TVTINGVKTNIPSGDTVVGTTAGTKNLEQLKN-----
WP_003097688.1	---DSTFDTSDLQAGD-----	ITKIE-----VTNLAEFGA-----
WP_033689127.1	---DTTISTSTLNAGS-----	ITKIE-----LTNLAEFGA-----
WP_020903468.1	---DTTISTSTLKAGD-----	VTKIE-----LTNLAEFGS-----
WP_000287308.1	---DTTISTSTLKAGS-----	ITKIE-----LTNLAEFGG-----
WP_033685380.1	---DVVFDTSTQAQNGD-----	TTTIE-----MKNLTGFGD-----
WP_033681547.1	---EISFDTSTRTQKGD-----	TTKIE-----LKNLAGFGD-----
WP_023948409.1	---EISFDTSTRTQKGD-----	TTKIE-----LKNLAGFGD-----
WT-GB1-YSIRK	-----EG-----	
WT-GA-YSIRK	-----	
YP_005861639.1	-----KPTPMQITQPTVKDITWTINGV-----	
WP_021874218.1	R-----NNFKNSNIAGNSVNIENGASLTINSSEINDGINLTDGGTVRVGDNA	
KFL97016.1	SWASGSSNQNGWTVKGWNYANFSGSKVNVAADANLTINRSAIGDGIHLANNGTVNVADGG	
KFL95230.1	SWASGSSNQNGWTVKGWNYANFSGSKVNVAADANLTINRSAIGDGIHLANNGTVNVADGG	
WP_033688740.1	-----NINWFDFAAASITYSNGTVVGPARKLAQPITKTITYPNGDRVEGK	
WP_003097688.1	-----AFP---VGKEITAADGTVIGKVKSIDY-----	
WP_033689127.1	-----GLA---VNTEIRATDGTVVGKVKSIDF-----	
WP_020903468.1	-----GLA---VNSEIRATDGTVVGKVKSIDF-----	
WP_000287308.1	-----GLA---VNTEIRETDGTVVGKVKSIDY-----	
WP_033685380.1	-----SFK---PGTKITAADGTVIGEVSSKET-----	
WP_033681547.1	-----AFK---AGTPIKAADGTTIGVVKSAAQ-----	
WP_023948409.1	-----AFK---VGTPIKAADGTTIGVVKSAAQ-----	

Figure A1. Continued.

WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	-----
WP_021874218.1	TL-----
KFL97016.1	QLTI-----
KFL95230.1	QLTI-----
WP_033688740.1	ITMVRDVTYADGSKGLTTDDKFLNSGTAKYVEHLYYTGGAQNDHELYEALQEGMKFNVKT
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----
WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	-----
WP_021874218.1	-----
KFL97016.1	-----
KFL95230.1	-----
WP_033688740.1	KVEGYALTATVIKLGSKAVDSDPNKTPAGPVNAVRDYGDWGSDKQLAKAQAEDKRFRKANA
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----
WT-GB1-YSIRK	---ETSVDSE-----
WT-GA-YSIRK	-----
YP_005861639.1	---EQTSAEF-----
WP_021874218.1	---NVNLTNA-----
KFL97016.1	---NMNTNND-----
KFL95230.1	---NMNTNND-----
WP_033688740.1	TKNKTTLANNPTVSVNGVALTLPEYPTETFANTDAYYSKAAAYNGAVKAYNQIDAQQLS
WP_003097688.1	---KTSTGNN-----
WP_033689127.1	---KTTTGNN-----
WP_020903468.1	---KTTTGNN-----
WP_000287308.1	---KTTTGNN-----
WP_033685380.1	---TNSVGSR-----
WP_033681547.1	---SNSTGNK-----
WP_023948409.1	---SNSTGNK-----

Figure A1. Continued.

WT-GB1-YSIRK	-----Y-----
WT-GA-YSIRK	-----
YP_005861639.1	-----HIDVNPV-----WNPKFSL-----
WP_021874218.1	-----STTATRYHVAGVFAKNGGN-----FISGYKSNV-----
KFL97016.1	-----LNTTARYHNAGIFAVGNGN-----FTTGYKSVV-----
KFL95230.1	-----LNTTARYHNAGIFAVGNGN-----FTTGYKSVV-----
WP_033688740.1	ATKLSETGFTYVNVNGKAVPNGSANLYKKIGNERKADVLITAQDTQWSYLRKAGLPTINF
WP_003097688.1	-----NNKSVPY-----WAQRTQRM-----
WP_033689127.1	-----NNKSVPY-----WAQRTQRM-----
WP_020903468.1	-----NNKSVPY-----WAQRTQRM-----
WP_000287308.1	-----NNKSTPY-----WGQRTQRM-----
WP_033685380.1	-----GGES-PF-----WSQRKKDCK-----
WP_033681547.1	-----GTTS-PY-----WAQRMKDGM-----
WP_023948409.1	-----GTTS-PY-----WAQRMKDGM-----

WT-GB1-YSIRK	-----DFNEP-----
WT-GA-YSIRK	-----S-----
YP_005861639.1	-----KPNPNSTDQNALKKMI-----
WP_021874218.1	-----NFNTGLGQAIAIGATR-----PTGTDSDFR-----
KFL97016.1	-----TLNTSIGQGIAMTGMR-----PYVTDTDVF-----
KFL95230.1	-----TLNTSIGQGIAMTGMR-----PYVTDTDVF-----
WP_033688740.1	DGSEMLTAFTSSRDESNVGVTFALSATYNGRIVDVNVIAADAEAGRTEIVQFETDGTKW
WP_003097688.1	-----TYDERVAEQPAIANET-----
WP_033689127.1	-----TYDQRVAEQPAVANET-----
WP_020903468.1	-----TYDQRVAEQPAVANET-----
WP_000287308.1	-----TYDQRVAEQPAVANET-----
WP_033685380.1	-----TYEERLAEQPAIPNEV-----
WP_033681547.1	-----TYEERLAEQPAIPNEV-----
WP_023948409.1	-----TYEERLAEQPAIPNEV-----

WT-GB1-YSIRK	-----EIDWSCNDEAVAEEE-----
WT-GA-YSIRK	-----
YP_005861639.1	-----PNY-----ESIYQLAINETNGVA
WP_021874218.1	GGYGARSRRNDGPTLVQLGDSSTFNFTGRDGIILGNANFISGENSNVHFENKGRGVALDL
KFL97016.1	GGYSARDRGDGSQINLGQYSTLNFTGRDGVILGNNSNFNVGDSANVHFENKGRGVALDL
KFL95230.1	GGYSARDRGDGSQINLGQYSTLNFTGRDGVILGNNSNFNVGDSANVHFENKGRGVALDL
WP_033688740.1	EQFMALNLQKDIDEKTAQGVYPSRQSDVNADGRLAEGYNPKDWNVDENGNP SAYGTNT
WP_003097688.1	-----GTYTY-----NIEWNEKAKDYPNIS
WP_033689127.1	-----GTYTY-----NIEWNDKVKDYPNVS
WP_020903468.1	-----GTYTY-----NIEWNDKVKDYPNVS
WP_000287308.1	-----GTYTY-----NIEWNDKVKDYPNVS
WP_033685380.1	-----GTTTY-----TIRWNDKAKNYAVTT
WP_033681547.1	-----GTITY-----TIEWNEKASNPVTT
WP_023948409.1	-----GTITY-----TIEWNEKASNPVTT

Figure A1. Continued.

WT-GB1-YSIRK	AAQEEVYTL -----
WT-GA-YSIRK	-----
YP_005861639.1	PGQDYQNSL-----
WP_021874218.1	AANSNIEISKHSTTYFHSVGKGTGSGSYD-----
KFL97016.1	AANSNINIDDHAVTYFHSVGKTTTNALGNTVGASGSFS-----
KFL95230.1	AANSNINIDDHAVTYFHSVGKTTTNALGNTVGASGSFS-----
WP_033688740.1	FGANYTSMGSKNSLPIALSQNVKTLMSMYLNSAGAQAQGTIGFMIYDGGDAPQSYGSAQHII
WP_003097688.1	FGAENLSGG-----
WP_033689127.1	FGASNLSGN-----
WP_020903468.1	FGASNLSGD-----
WP_000287308.1	FGASNLSGS-----
WP_033685380.1	FYAENLTAI-----
WP_033681547.1	YSVENLTGS-----
WP_023948409.1	YSVENLTGS-----
 WT-GB1-YSIRK	 NYYAQRT -----
WT-GA-YSIRK	DYY -----
YP_005861639.1	PYPSSKINSA-----VNYGTVITI-----
WP_021874218.1	GYNYIGVNEG-----GNITVD-----
KFL97016.1	GYNYIGVNEG-----GNITVG-----
KFL95230.1	GYNYIGVNEG-----GNITVG-----
WP_033688740.1	GDFNKEVVVDGVTQVTATQPYLGNVKGDPDFRSTKTDPSGGWVLDLITSEKYKETPLE
WP_003097688.1	GYLAPQISKD-----TDYKATIKID-----
WP_033689127.1	GYLAPQISKD-----TPYTATIKID-----
WP_020903468.1	GYLAPEISKD-----TPYTATIKID-----
WP_000287308.1	GYAPPEISKD-----TPYTATIKID-----
WP_033685380.1	DYYAPNISKD-----TEYTAAISVN-----
WP_033681547.1	GYAPQISKD-----TEYTAEIKVD-----
WP_023948409.1	GYAPQISKD-----TEYTAEIKVD-----
 WT-GB1-YSIRK	 -----
WT-GA-YSIRK	-----
YP_005861639.1	-----
WP_021874218.1	-----EYATFRVILEGRGDNP-----WDD-----
KFL97016.1	-----KFATFRVILEGRGNNN-----YDD-----
KFL95230.1	-----KFATFRVILEGRGNNN-----YDD-----
WP_033688740.1	SGKTVVTTDKGVTGKYLLLPNGNAVIEKSDNTRVLLNQGDVIEMVNPSTKLPIRGIYNHT
WP_003097688.1	-----GRTVVEHT-----YTR-----
WP_033689127.1	-----GRTVLEHT-----YTR-----
WP_020903468.1	-----GRTVLEHT-----YTR-----
WP_000287308.1	-----GRTVLEHT-----YTR-----
WP_033685380.1	-----GQPILEHK-----YTH-----
WP_033681547.1	-----GQKVLEHT-----YTH-----
WP_023948409.1	-----GQKVLEHT-----YIH-----

Figure A1. Continued.

WT-GB1-YSIRK	-----QGQNGATTVKASSPR-----
WT-GA-YSIRK	-----
YP_005861639.1	-----PMPKGYVLDESATMQLNNFG-----
WP_021874218.1	-----VVSLDSQNANTTAAFTSKKGAIVDIRDDN-----
KFL97016.1	-----VVSLDSQNTNTNAAFTSKTGAIVDIRDDN-----
KFL95230.1	-----VVSLDSQNTNTNAAFTSKTGAIVDIRDDN-----
WP_033688740.1	TGALGEGTLGDEGESQLLDPAVATEYKLRQAQGNEYVLDGVRANLGVNNDKAYVRGWVDF
WP_003097688.1	-----KGQQATFSKQGTSSASLRTNN-----
WP_033689127.1	-----KGQQPSYKQGTSSASLSENN-----
WP_020903468.1	-----KGQKANYKQGTSSASLSENN-----
WP_000287308.1	-----KGQPPNYKQGTSSASLSENN-----
WP_033685380.1	-----KASKSAAQKQNTSATIMGDN-----
WP_033681547.1	-----KATKPSVQKQSTSVSLLTDN-----
WP_023948409.1	-----KATKPSVQKQSTSVSLLTDN-----

WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	-----
WP_021874218.1	-----
KFL97016.1	-----
KFL95230.1	-----
WP_033688740.1	NNNGKFDLNESEIVEVNQNGTYSIKFKNTPQLLDTSDSLGVRLRISLTKDEILEPTGV
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----

WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	-----DKTAITQDGNIIITVPKGSGTQNWNSGGPY-----
WP_021874218.1	-----TNFYAELISFPLG--GSNSRIDIQDPLMLNLQ-----RY-----
KFL97016.1	-----TNFYAELISFPLG--ASNTRIDIHDPLMLNLQ-----RY-----
KFL95230.1	-----TNFYAELISFPLG--ASNTRIDIHDPLMLNLQ-----RY-----
WP_033688740.1	ASSGEVEDFETHVIHMPRGTKHETKDFQGREQTVKLPNTAMFTASGKNKDSNYQWAQIE
WP_003097688.1	-----GLGYVSNELKSKSDTLEINTDS-----DIRY-----
WP_033689127.1	-----GLTYLNNEQTGRSDSIVLKTDS-----DIRY-----
WP_020903468.1	-----GLKYLNEQISRSDSIVLKTDS-----DVR-----
WP_000287308.1	-----GLTYLNNEQTGRSDSIVLKTDS-----DVR-----
WP_033685380.1	-----ILGYKGSEVRKSDAVVINTDS-----DVR-----
WP_033681547.1	-----GVNYNGSSLVKKNDVVINTDT-----DVR-----
WP_023948409.1	-----GVNYNGSSLVKKNDVVINTDT-----DVR-----

Figure A1. Continued.

WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	-----
WP_021874218.1	-----
KFL97016.1	-----
KFL95230.1	-----
WP_033688740.1	NDNLPPKIVLTDKQVASEEAYTPTDSELP SNYRK GDDGKV FVERNGATVFTGEYVTVKDK
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----
 WT-GB1-YSIRK	 -----
WT-GA-YSIRK	-----
YP_005861639.1	-----QLVGSYNIPMPNTATTYTAD-----
WP_021874218.1	-----SKGGATTGWMPITGGDMINTTSAEYTSNLIYMSGNKG VFSVSGGDYDPSNPNSSGF
KFL97016.1	-----SSGGPTTGWMPIGGDMINTT SNQYTANLIYMSGSKGVFSVDGTNYVVYQKIKSDG
KFL95230.1	-----SSGGPTTGWMPIGGDMINTT SNQYTANLIYMSGSKGVFSVDGTNYVVYQKIKSDG
WP_033688740.1	NNKVLGKGLKITNPLNNKTEYLLDTYTEYDTAGNEVGTYKVN PASNGKNV SIGNGLYETT
WP_003097688.1	-----GVGSKFTIKLPN-----SEFTEF-----
WP_033689127.1	-----GVGSKFTIKLPN-----ADFTEF-----
WP_020903468.1	-----GVGSKFTIKLPN-----ADFTEF-----
WP_000287308.1	-----GVGSKFTIKLPN-----ADFTEF-----
WP_033685380.1	-----GKGSKFTITLPN-----DDFTYF-----
WP_033681547.1	-----GKGSKFTIDLPS-----DEFTYF-----
WP_023948409.1	-----SKGSKFTIDLPN-----DEFTYF-----
 WT-GB1-YSIRK	 ----- EALEYFQAFI
WT-GA-YSIRK	----- KRLVNSAKTV
YP_005861639.1	-----APITIVQKLN
WP_021874218.1	VVYQRIKSDGSKQIW-----LNVNDVNIPM
KFL97016.1	SKQIW-----LNVNGVNIPM
KFL95230.1	SKQIW-----LNVNGVNIPM
WP_033688740.1	LTFKPVDAYVGTAKGIAVRAWDDNNSSTGWEATNDTIETSKTSTTLAEKDKVLENTNNGN
WP_003097688.1	-----KELEGSTNFV
WP_033689127.1	-----KELEGSSNFV
WP_020903468.1	-----KELEGSSNFV
WP_000287308.1	-----KELEGSSNFV
WP_033685380.1	-----KAIDGSKN--
WP_033681547.1	-----RELNGGKN--
WP_023948409.1	-----RELNGGKN--

Figure A1. Continued.

WT-GB1-YSIRK	NE-----
WT-GA-YSIRK	AG-----
YP_005861639.1	ND-----
WP_021874218.1	NG-----
KFL97016.1	SG-----
KFL95230.1	SG-----
WP_033688740.1	NGYKSMDSYIPTVIDVRPVGEDTITEDVQGKPQSSNPTIPAYATVETVTNDKIEDTKYA
WP_003097688.1	NG-----
WP_033689127.1	NG-----
WP_020903468.1	NG-----
WP_000287308.1	NG-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----

WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	-----GSQTKTWTGPTVSQDFYGAN-----
WP_021874218.1	-----FQTKDIWNNQANPDVSITGNGLT-----
KFL97016.1	-----FQTKDIWDNQANPDVSIKGNDLT-----
KFL95230.1	-----FQTKDIWDNQANPDVSIKGNDLT-----
WP_033688740.1	ANFVILDKTKKPTLATQKENPGKLYTEDTKVDKETTLTLTDGTTATYKPVDKIPANTVIA
WP_003097688.1	-----LNTASTLNPKNKGSITYRLANRW-----
WP_033689127.1	-----LNTASTINPNKGSITYRPASRW-----
WP_020903468.1	-----LNTASTINPNKGSITYRPASRW-----
WP_000287308.1	-----LNTASTINPNKGSITYRPASRW-----
WP_033685380.1	-----TVTNTNKADSISYRPAGRW-----
WP_033681547.1	-----SI-----KSSEVTYRPSNRW-----
WP_023948409.1	-----SI-----KSSEVTYRPSNRW-----

WT-GB1-YSIRK	--NGLDAADFN-----
WT-GA-YSIRK	-----
YP_005861639.1	--DQIPLAQVPLYAKAA-----
WP_021874218.1	--GGIRANQVHNY---NGSPLTGKDAPYYGISTQRASQQIWIPHRTPLEITGNHTNTIKY
KFL97016.1	--SGIRANQVHNY---DGTPLTGKDAPYYGISTQRASQQIWFPKTMQEVVGSHTNTIKY
KFL95230.1	--SGIRANQVHNY---DGTPLTGKDAPYYGISTQRASQQIWFPKTMQEVVGSHTNTIKY
WP_033688740.1	KDGNVEVTGTGNVVLNNVRLVTGSQIPA-----GSKPQSNHPTTVEVQVTLADGTEQTI
WP_003097688.1	--ANVKANENNVWILQD----GREGGF-----TLTPKLISPTLELVTVT-----
WP_033689127.1	--ANVRANENNVWILQD----GRDSGF-----TLTPRLISPTLELVTVT-----
WP_020903468.1	--ANVKANENNVWILND----GRDSGF-----TLTPRLISPTLELVTVT-----
WP_000287308.1	--ANVKANENNVWILND----GRDSGF-----TLTPRLISPTLELVTVT-----
WP_033685380.1	--NNVQANANNVWILND----GRDTNF-----TLKATIKSPTELEIEVI-----
WP_033681547.1	--SNVQANQNNVWILQD----GRDTSF-----TLKATLKSPTRLELEVI-----
WP_023948409.1	--SNVQANQNNVWILQD----GRDTSF-----TLKATLKSPTRLELEVI-----

Figure A1. Continued.

```

WT-GB1-YSIRK -----
WT-GA-YSIRK -----
YP_005861639.1 --YGGNQLLNNGHKQIVAYFGATNESIASYSNYSSNFTFNFDSESLGVTELKTPPTIPGT--
WP_021874218.1 VDEQGNEIFPENTS-----SLNLKRNIIIDITQDIKKIQDYALNHTADETLEYIKNSQ
KFL97016.1 VYEDGTPVLDENGNQIVKTQNLNLTRKLTLDITDDKIEEIQKYALHNADQTLEYIKNAQ
KFL95230.1 VYEDGTPVLDENGNQIVKTQNLNLTRKLTLDITDDKIEEIQKYALHNADQTLEYIKNAQ
WP_033688740.1 PAGG---TIPGGATIKTFTN-----TATNTFNNVTYNN-GQTIPAA--SAGKIS
WP_003097688.1 --EG---YIQEGSNISLSLQSLGVEKVIKDKTLTSEFSNVTYNEFGIITGG---TAGN--
WP_033689127.1 --EG---AIQEGSTVSMPLQSLGIEKVIKDKTLTSEYSNITYEN-GLIKQG---YVGN--
WP_020903468.1 --EG---AIQEGSVVSMPLQSLGIEKVVKDKTLTSEYANITYEN-GLIKEG---YVGN--
WP_000287308.1 --EG---AIQEGSVVSMPLQSLGIEKVIKDKTLTSEYSKITYEN-GLIKEG---YVGN--
WP_033685380.1 --DG---AIQEDSIVSIGLDKLGVEKAITNRTFSDEYSKFIYDEAGRLKDG--VVGD--
WP_033681547.1 --DG---VIQEGSTVSMALDKLGIEKVLTPRTFSDDFSKIKYDELGRVAKGS--TVGN--
WP_023948409.1 --DG---VIQEGSTVSMALDKLGIEKVLTPRTFSDDFSKIKYDELGRVAKGS--TVGN--

WT-GB1-YSIRK -----WSYDSESRAF-----
WT-GA-YSIRK -----
YP_005861639.1 ---SNKYTITYADGTT--SEGQVNAGNTITGTGVI-----TNIVVSPDNF-----
WP_021874218.1 SVAQDSGWKFTNGSGQT--VTDPYATVESPKLDG-----YTATIQTSTNVQGLKVG
KFL97016.1 GVSEDSGWVYTDAQNT--VTDPYATVVSPVEDG-----YTASIESSNVPGITG
KFL95230.1 GVSEDSGWVYTDAQNT--VTDPYATVVSPVEDG-----YTASIESSNVPGITG
WP_033688740.1 SLAKETSADQLVEKGKSITLDGTTYNNNDVIPKGRTRMTTYEDLRNVTLPNAVHIDPQTG
WP_003097688.1 ---DKTAATLTVSGGTS--INGEKEDVTTTVANG-----WEINASGTPLGEPPTG
WP_033689127.1 ---DKTAATLTVSGGES--VNGEKEDVITKVPNG-----WSIVGDGRVQGEPTG
WP_020903468.1 ---DKTAATLTVSGGES--VNGEKEDVATTVPNG-----WSVVGDKGVQGEPTG
WP_000287308.1 ---DKTAATLTVSGGES--VNGEKEDVATTVPNG-----WSVVGDKGVQGEPTG
WP_033685380.1 --TDKTAATLTVSGGTF--LNGGEENVTTTKVANG-----WKVEVGAGGNSQFETG
WP_033681547.1 ---DKTSATLSVTGGTF--LNGQTEEIATKVNNG-----WEVSVIPNGSTTVETG
WP_023948409.1 ---DKTSATLSVTGGTF--LNGQTEEIATKVNNG-----WEVSVIPNGSTTVETG

WT-GB1-YSIRK -----
WT-GA-YSIRK -----
YP_005861639.1 -----ERDQSTAVDLPTNKFAN-----
WP_021874218.1 EDASSVTAKFAVNPSIEDIVQNGELTD--SYKNDGITGIPDNYVTV-----
KFL97016.1 ADGTSVTAKLQYKEELVQNGELSN---NYKQNGLSAILPDNYET-----
KFL95230.1 ADGTSVTAKLQYKEELVQNGELSN---NYKQNGLSAILPDNYET-----
WP_033688740.1 EVTSVPRRYTKVTETEIVIENEGTYTLNQDTGEITFIPDPKFVGTGTGVTKQPDVDYND
WP_003097688.1 AVVIGLK-----DLETGKIIGHATTKYDG-----
WP_033689127.1 AVVRTFK-----DLVTGEVIGFEPTRYTG-----
WP_020903468.1 AVVRTFK-----DLVTGEVIGFEPTRYTG-----
WP_000287308.1 AVVRTFK-----DLVTGEVIGFEPTRYTG-----
WP_033685380.1 AVTITLV-----NIETGEEFAYEPTSVDG-----
WP_033681547.1 AVTLTLK-----DLETGKIIGYIPTEYNG-----
WP_023948409.1 AVTLTLK-----DLETGKIIGYIPTEYNG-----

```

Figure A1. Continued.

```

WT-GB1-YSIRK -----
WT-GA-YSIRK -----
YP_005861639.1 -----
WP_021874218.1 -----VVYKKAKEKGSVKVYHDDTTNTEIPNT
KFL97016.1 -----VVYKKAKEVTNTLKFYDDTTKSYISTV
KFL95230.1 -----VVYKKAKEVTNTLKFYDDTTKSYISTV
WP_033688740.1 KVAGDPVTSKYGTDYGKAKYIPIVKPQSKASITRTIHVYENANDNPTSQDSYKDNDPIL
WP_003097688.1 -----NRPLNEDGTDKDNNTNL
WP_033689127.1 -----NIPLSEDGSKDYTNVL
WP_020903468.1 -----NIPLSEDGSKDYTNVL
WP_000287308.1 -----NIPLSEDGSKDYTNVL
WP_033685380.1 -----YAKPNEDGSYETKNIL
WP_033681547.1 -----YAPLKEDGTYETKNIL
WP_023948409.1 -----YAPLKEDGTYETKNIL

```

```

WT-GB1-YSIRK -----
WT-GA-YSIRK -----
YP_005861639.1 -----
WP_021874218.1 EYNTGSVDA-----GTKVDYTTT-----
KFL97016.1 ADQTAT-----GKENDDVNFKDGA-----
KFL95230.1 ADQTAT-----GKENDDVNFKDGA-----
WP_033688740.1 AIDNTPVTRTQTIDYTRDYKIFSEAGTTDTAITTTNQVTDASGNIYNVGDTIPAGTQFNQ
WP_003097688.1 -----GKKYDVSDYIP-----
WP_033689127.1 -----GNKYDVSNDNV-----
WP_020903468.1 -----GNKYDVSNDPV-----
WP_000287308.1 -----GNKYDVSNDPV-----
WP_033685380.1 -----GKKYDVSNQVP-----
WP_033681547.1 -----GKKYDVTNQIP-----
WP_023948409.1 -----GKKYDVTNQIP-----

```

```

WT-GB1-YSIRK -----
WT-GA-YSIRK -----
YP_005861639.1 -----
WP_021874218.1 -----
KFL97016.1 -----
KFL95230.1 -----
WP_033688740.1 GSIIIGKWTASSDQNSKFKEIISPTVKGYTAEVVTADFTPRADGKMGHIHNGKQPVGLYT
WP_003097688.1 -----
WP_033689127.1 -----
WP_020903468.1 -----
WP_000287308.1 -----
WP_033685380.1 -----
WP_033681547.1 -----
WP_023948409.1 -----

```

Figure A1. Continued.

```

WT-GB1-YSIRK      -----TASEKIEGEA-----
WT-GA-YSIRK      -----VKKLQAQV-----
YP_005861639.1    -----QTTQSVNAFE-----AYGAVPETV
WP_021874218.1    -----TTITNLENQG-----YVY-----VSTDGTIPSTI
KFL97016.1        -----STVKSLEDQG-----YKF-----INVTDGTPDDTNATVL
KFL95230.1        -----STVKSLEDQG-----YKF-----INVTDGTPDDTNATVL
WP_033688740.1    PVADNNKDVGAYEPLVSEVRSDDKDDFDMYVVYKADIQKAKVTYIDLDATGDARILEVQN
WP_003097688.1    -----PLIKEVDGEE-----YIL-----ANIPTTGIEGTL SVTN
WP_033689127.1    -----DLVKEVNGEE-----YIL-----ADIPVENARGT LSVTK
WP_020903468.1    -----DLVKEVNGEE-----YIL-----ADLPAENAKGT LSVTK
WP_000287308.1    -----DLVKEVNGEE-----YIL-----ADIPAENTKGT LSVTK
WP_033685380.1    -----DLTKTIKGV-----YIR-----VDVPSKGT TGTVNIGP
WP_033681547.1    -----DLVKPIAGVD-----YIL-----AEIPPKGPAGT INVSD
WP_023948409.1    -----DLVKPIAGVD-----YIL-----AEIPPKGPAGT INVSD

WT-GB1-YSIRK      -----
WT-GA-YSIRK      -----
YP_005861639.1    KS-----GTQLVANLFTFTGTIQQGN-----
WP_021874218.1    EG-----NQNVVTVHMK-HGVQP-----
KFL97016.1        SG-----DTFSDVDFGKFGKDGKTFVVHLTHK-----
KFL95230.1        SG-----DTFSDVDFGKFGKDGKTFVVHLTHK-----
WP_033688740.1    ANPAPATGADAKTTYGVATLQKSHTAIPYLTAETIKKYEDRGYELVTD DYTNNQTGTAI
WP_003097688.1    TR-----ARDLYSTEELKENGLNASAY-----
WP_033689127.1    TR-----ARDLYSEEELKAKGINGSAF-----
WP_020903468.1    TR-----ARDLYSEEELKAKGINGSAF-----
WP_000287308.1    TR-----ARDLYSEEELKAKGINGSAF-----
WP_033685380.1    KR-----ASAIYSSEELQANGVNPNAF-----
WP_033681547.1    KR-----IRDIYTADEITAAGLNPNAY-----
WP_023948409.1    KR-----IRDIYTADEITAAGLNPNAY-----

WT-GB1-YSIRK      -----SVDSEYD-----FIRPDFD-----
WT-GA-YSIRK      -----VE-----SAKKARV-----
YP_005861639.1    -----ITKYLTS-----KIQVDQT-----VVSSADLTSS-----S
WP_021874218.1    -----VTPDTPTPDVPKNTPAEAQPDQL-----TKKVNLTVNYVNS--D--G
KFL97016.1        -----VVPVTPDTPNVPSNSKVS KDDLT-----KTATRTIHYVE-----N
KFL95230.1        -----VVPVTPDTPNVPSNSKVS KDDLT-----KTATRTIHYVE-----N
WP_033688740.1    EGGRKFDDDKQAFNV-----YLRHKKVTRKIKDTQEVTRTIEYKYASTDDVPA
WP_003097688.1    -----VTPVDYN-----YVKKTRV-----EEVNRTIKFYVYA--DDA-Q
WP_033689127.1    -----VNPVNYD-----YVKKTKV-----EEVNRTIKFYVYA--DNV-E
WP_020903468.1    -----VTPAEYD-----YVKKTKV-----EEVNRTIKFYVYA--DNV-A
WP_000287308.1    -----VTPAEYD-----YVKKTKV-----EEVNRTIKFYVYA--DNV-A
WP_033685380.1    -----VNNVVYY-----YVKKTKV-----EEVNRTIKFYVYA--DDV-K
WP_033681547.1    -----VNNVIYS-----YVKKTKV-----EEVNRTIKFYVYA--DDV-K
WP_023948409.1    -----VNNVIYS-----YVKKTKV-----EEVNRTIKFYVYA--DDV-K

```

Figure A1. Continued.

WT-GB1-YSIRK	-----
WT-GA-YSIRK	SEAT -----
YP_005861639.1	GIFGYQSNTATGQSNVGYLSV-----YAGGGQTNNIY-----EPIF
WP_021874218.1	STFTATVPANAKQTVTFTGTAYVDKVTGQLVNATQ-QNGQWVIDENN-----TATPQI
KFL97016.1	DQNGAELKESTVQTVNYTGTAYVDVVTGQMVNAK--ADGKDAQGNTTYVVDTDNKKQPSI
KFL95230.1	DQNGAELKESTVQTVNYTGTAYVDVVTGQMVNAK--ADGKDAQGNTTYVVDTDNKKQPSI
WP_033688740.1	DKRGTTAAPTVTETLHFERDRTIDYTLAAKEYPTEYAAYKAVLDASGYDSPEEYKARVVY
WP_003097688.1	GLAGQQVFDPTTQTVSYKGTVKTN-----AEGKAEIGSND-----KPIY
WP_033689127.1	GLAGTEVFPSQKQTVSYTGSIKLT-----AEGKAVINSND-----RPVY
WP_020903468.1	GLAGTEVFPSQKQTVSYTGSIKLT-----AEGKAVINSND-----RPVY
WP_000287308.1	GLAGTEVFPSQKQTVSYTGSIKLT-----AEGKAVINSND-----RPVY
WP_033685380.1	DLAGQEVFEPKQTVSYTGTIQVN-----DKNEAQVDANR-----KPIY
WP_033681547.1	NLAGQQVFEPKQTVSYTGTVKLN-----SDGKAAVDSND-----KPIY
WP_023948409.1	NLAGQQVFEPKQTVSYTGTVKLN-----SDGKAAVDSND-----KPIY

WT-GB1-YSIRK	----- W -----
WT-GA-YSIRK	-----
YP_005861639.1	YYVLPE-----W-----
WP_021874218.1	T-----W--TSDKTSFD-----KVVSP--
KFL97016.1	TWTTDNNGKFAQVTPDASIKKGGDTWT--TGVKSVDEKNAPDVLITIGKTTNEDVYVPYT
KFL95230.1	TWTTDNNGKFAQVTPDASIKKGGDTWT--TGVKSVDEKNAPDVSTITGKTTNEDVYVPYT
WP_033688740.1	YDHITNKAIADATDAQKAIIVTFGPWTP-VGGTSNDAITLSDAEKAKDDKFNL-VNSP--
WP_003097688.1	IN-----WKG-IG--DTN-----LPE-VTVP--
WP_033689127.1	IN-----WKG-TDGQSTD-----LPE-LAVP--
WP_020903468.1	IN-----WKG-TDGQSTD-----LP-ELAVP--
WP_000287308.1	IN-----WKG-TDGQSTD-----LPE-LAVP--
WP_033685380.1	IN-----WVG-NG--DTN-----LPE-VTVP--
WP_033681547.1	VN-----WVG-NG--DTN-----LPE-VTVP--
WP_023948409.1	VN-----WVG-NG--DTN-----LPE-VTVP--

WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	-----FSVYDFSTDYTKL--
WP_021874218.1	-----VEQNY-HLISISDHQDGNVATITGLTKDSGDITVTVTYAPNGK
KFL97016.1	LSQKTYTGKETKTVTRVINYLNETKQPVSDAVEQTTLSRTQIKDEKGNVIGYGTVSE
KFL95230.1	LSQKTYTGKETKTVTRVINYLNETKQPVSDAVEQTTLSRTQIKDEKGNVIGYGTVSE
WP_033688740.1	-----EVTGYVP--D-----NATVEATAAIDAEADDYKITVLYTPVAQ
WP_003097688.1	-----QKEGYIASVE-----KVPVQPTTATD---EDYEYVVVTSPI-Q
WP_033689127.1	-----QKEGYIASVE-----KVPVQATTATD---EDYEYVVVYTAI-Q
WP_020903468.1	-----QKEGYIASVE-----KVPVQATTATD---EDYEYVVVYTAI-Q
WP_000287308.1	-----QKEGYIASVE-----KVPVQATTATD---EDYEYVVVYTAI-Q
WP_033685380.1	-----QKEGYIASVE-----KVPVQPTTATD---EDYEYVVVYTAI-Q
WP_033681547.1	-----QKEGYIASVE-----KVPVQPTTATD---EDYEYVVVTSPI-Q
WP_023948409.1	-----QKEGYIASVE-----KVPVQPTTATD---EDYEYVVVTSPI-Q

Figure A1. Continued.

WT-GB1-YSIRK	----- SGLE ----
WT-GA-YSIRK	----- DGLS ----
YP_005861639.1	-----PDFV-PN
WP_021874218.1	IIPVDPSGNPIPDAPTFQYPTDPTDPSKVTPNEPVPNPVPGYTPSVPTVTPID-PGKDTPV
KFL97016.1	DGHSYTLNNDWTIDKNGWVAQVSPDETAKGYKETPHFE-----DGKDAST
KFL95230.1	DGHSYTLNNDWTIDKNGWVAQVSPDETAKGYKETPHFE-----DGKDAST
WP_033688740.1	-----KAVVKFVEVDPTNTDKVI-----TPGLADPI
WP_003097688.1	-----KAKTTFVYQDKDGNVKQV-----EGNT-PI
WP_033689127.1	-----KAKTTFV--DEKGNA--I-----PGVA-EI
WP_020903468.1	-----KAKTTFV--DEKGNAI-----PGVA-EI
WP_000287308.1	-----KAKTIFV--DEKGNA--I-----PGVA-EI
WP_033685380.1	-----KAKTTFV--DEKGN--PI-----PGVD-EI
WP_033681547.1	-----KAKTTFVYQDKDGNVKQV-----EGNT-PI
WP_023948409.1	-----KAKTTFVYQDKDGNVKQV-----EGNT-PI

WT-GB1-YSIRK	----- EAEDD -----
WT-GA-YSIRK	-----
YP_005861639.1	TNNG-----VIGT--PKLSV-----
WP_021874218.1	PYTPETPAKDQKAVVNYVDAEDNKLITSS-----
KFL97016.1	VAADTPSVTDPQDVTNVFYDHDTPVTPDKPGHGLTHDDLNKDVRTTINYVDTTGAAVN
KFL95230.1	VAADTPSVTDPQDVTNVFYDHDTPVTPDKPGHGLTHDDLNKDVRTTINYVDTTGAAVN
WP_033688740.1	AVTG-----KSEAAYPATTA-----
WP_003097688.1	SETG-----KGGD--KLTKA-----
WP_033689127.1	TEQG-----GSET--PLTKE-----
WP_020903468.1	TEQG-----GSET--PLTKE-----
WP_000287308.1	TEQG-----GSET--PLTKE-----
WP_033685380.1	TEQG-----GSEA--PLTKE-----
WP_033681547.1	SETG-----KGGD--KLTKA-----
WP_023948409.1	SETG-----KGGD--KLTKA-----

WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	-----
WP_021874218.1	-----
KFL97016.1	GAPDGKSTYTQTAHFTRTAIVDKVNDKLLGYDINGDGSVDISPAGDFAWKSTDANLPAV
KFL95230.1	GAPDGKSTYTQTAHFTRTAIVDKVNDKLLGYDINGDGSVDISPAGDFAWKSTDANLPAV
WP_033688740.1	-----
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----

Figure A1. Continued.

WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	-----
WP_021874218.1	-----
KFL97016.1	TSKAPSEVGYDSVDTPVVQATTVAYNSEPINVTVTYSKNAQQGSFQIHYIDEDNNNAILH
KFL95230.1	TSKAPSEVGYDSVDTPVVQATTVAYNSEPINVTVTYSKNAQQGSFQIHYIDEDNNNAILH
WP_033688740.1	-----
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----

WT-GB1-YSIRK	-EEVQEE-----IYTF-----VYII-----
WT-GA-YSIRK	-----GF-----
YP_005861639.1	-FTVPTE-----ISGLSRQVVKIDYSGTGYNFLAGQGANN-----
WP_021874218.1	-GDLTGKAGTKIDYSTNSTIEDLTNK-----GYVLVNDGFPKD-----
KFL97016.1	QDTVSDKIGDSVTYSTADQIQLWESK-----GYVLDQDGYTTQTTVNEDNNGKTYI
KFL95230.1	QDTVSDKIGDSVTYSTADQIQLWESK-----GYVLDQDGYTTQTTVNEDNNGKTYI
WP_033688740.1	-TSVTDK-----IAELVKK-----GYELVDNGFVSA-----
WP_003097688.1	-DEIAAK-----IKEAQNK-----GYELVSNTYPTD-----
WP_033689127.1	-ADV KAK-----IAELENK-----GYELVSNTYPEG-----
WP_020903468.1	-ADV KAK-----IKELENK-----GYELVSNTYPEG-----
WP_000287308.1	-ADV KAK-----IKELENK-----GYELVSNTYPEG-----
WP_033685380.1	-ADV KAK-----IKELENK-----GYELVSNTYPEG-----
WP_033681547.1	-DEIAAR-----IKEAQNK-----GYEVVSNTYPTD-----
WP_023948409.1	-DEIAAR-----IKEAQNK-----GYEVVSNTYPTD-----

WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	-----
WP_021874218.1	-----
KFL97016.1	VSFKHGRKNGTTETLVPTETIHFAQYADGTKAADDVHGNAGDFKFTRTPIIDTVTGQIVDP
KFL95230.1	VSFKHGRKNGTTETLVPTETIHFAQYADGTKAADDVHGNAGDFKFTRTPIIDTVTGQIVDP
WP_033688740.1	-----
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----

Figure A1. Continued.

WT-GB1-YSIRK	-----	QN
WT-GA-YSIRK	-----	LK
YP_005861639.1	QIHLNLTLPDGTNGTYQGIIYI-----	VS
WP_021874218.1	ATYDND--DNITTQYTVVLRHGTQ-----	TN
KFL97016.1	GTWNKESYTFDDGQKNVKVINGYVADKATYGKNTATPTDLNVEDTVTYRKISNIIPVDEN	
KFL95230.1	GTWNKESYTFDDGQKNVKVINGYVADKATYGKNTATPTDLNVEDTVTYRKISNIIPVDEN	
WP_033688740.1	DKFDKD--AAVDQEYVVKFKAKVVDVPSFDPTKPASN-----	DN
WP_003097688.1	GAFDKD--VNTDQEFTVTLKERVV-----	DQ
WP_033689127.1	GKFDTD--KDTDQEFKVLKQKEV-----	DQ
WP_020903468.1	GKFDTD--KDTDQEFKVLKQKEV-----	DQ
WP_000287308.1	GKFDTD--KDTDQEFKVLKQKEV-----	DQ
WP_033685380.1	GKFDE--AGVDQEFKVTLKERVV-----	DQ
WP_033681547.1	GVFDKD--VDTDQEFTVTLKERVV-----	DQ
WP_023948409.1	GVFDKD--VDTDQEFTVTLKERVV-----	DQ

WT-GB1-YSIRK	TKGKNGATTV----	KASSP-----	EEAKAYFEF	FAKENDLGEL-----
WT-GA-YSIRK	SQ-TPAEDTI-----			
YP_005861639.1	PTTKLTNTAY----	NPNNT-----	SNFAPSGITFNP	DWVQGNTSNLY-----
WP_021874218.1	PG-KPG-EPI----	NPNDP-----	DGPKYPTGS--	NEVTKTVTRTIQ-----
KFL97016.1	GNQIPGTTTPVDYKNDPSDPTKVT	PDEESPKVPSGWTIS	PNQPEGVTPNTTTNTAKVTPVD	
KFL95230.1	GNQIPGTTTPVDYKNDPSDPTKVT	PDEESPKVPSGWTIS	PNQPEGVTPNTTTNTAKVTPVD	
WP_033688740.1	PKPTPGVTPI----	DPNNP-----	DGPKWTEALINAVKVQEEVTRTIK	-----
WP_003097688.1	PK-TSG-TPV----	DPNNP-----	DGPKYPAGL-EEKDLNKTIVTRTIT	-----
WP_033689127.1	PK-TPG-TPV----	DPNNP-----	DGPKYPAGL-EEKDLNKTIVTRTIT	-----
WP_020903468.1	PK-TPG-TPV----	DPNNP-----	DGPKYPAGL-EEKDLNKTIVTRTIT	-----
WP_000287308.1	PK-TPG-TPV----	DLNNP-----	DGPKYPAGL-EEKDLNKTIVTRTIT	-----
WP_033685380.1	PK-TPG-TPV----	DPNNP-----	DGPKYPAGL-EEKDLNKTIVTRTIT	-----
WP_033681547.1	PK-TPG-TPV----	DPNNP-----	DGPKYPAGL-EEKDLNKTIVTRTIT	-----
WP_023948409.1	PK-TPG-TPV----	DPNNP-----	DGPKYPAGL-EEKDLNKTIVTRTIT	-----

WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	-----
WP_021874218.1	-----
KFL97016.1	PTKPTNVVYTKDNAPVDKATVIVRYHDDTTNLDLPESFDSGNKEVGTDGTGYTQADINKVV
KFL95230.1	PTKPTNVVYTKDNAPVDKATVIVRYHDDTTNLDLPESFDSGNKEVGTDGTGYTQADINKVV
WP_033688740.1	-----
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----

Figure A1. Continued.

```

WT-GB1-YSIRK -----
WT-GA-YSIRK -----
YP_005861639.1 -----YVGANGF-----
WP_021874218.1 -----YLDEDGN-----
KFL97016.1 QEYEAKGYYYVTTDGTLPPTIPAGGATIVVHLAHNQIPVGPDTPDKHGVDPPQVKKAYTS
KFL95230.1 QEYEAKGYYYVTTDGTLPPTIPAGGATIVVHLAHNQIPVGPDTPDKHGVDPPQVKKAYTS
WP_033688740.1 -----YVYEDGT-----
WP_003097688.1 -----YVYEDGT-----
WP_033689127.1 -----YVYADGT-----
WP_020903468.1 -----YVYEDGT-----
WP_000287308.1 -----YVYEDGT-----
WP_033685380.1 -----YVYEDGT-----
WP_033681547.1 -----YVYEDGT-----
WP_023948409.1 -----YVYEDGT-----

WT-GB1-YSIRK -----DWT---YDEDTKTFTAR
WT-GA-YSIRK -----KSLELSEAKTL
YP_005861639.1 TINQVGGANTASVAQGNQNDILVDSGVSDLY-----GSNEMEYAVR
WP_021874218.1 KVSQSV-----EQPVNFTASGVLDKVTGEWT---TPLTWS---VDQTVSAVKSP
KFL97016.1 TLHYQDSEGKTLSPDQQQTSTWTRTVTVDTVTNQIVNGGKYDTNWTLQDANDKYSNFTVP
KFL95230.1 TLHYQDSEGKTLSPDQQQTSTWTRTVTVDTVTNQIVNGGKYDTNWTLQDANDKYSNFTVP
WP_033688740.1 PVAESDLTSSVADKK--VKTLKFTRSGKINVATGEIT---YG-DWS---ADQTFEAVTSP
WP_003097688.1 PVLNED--GTPKTV--TQEAKFTREAKVNLVTGEVT---YG-DWT---PAQDLSEVTSP
WP_033689127.1 PVLNED--GTPKTV--TQEAKFTREAKVNLVTGDVT---YG-DWS---EAKDLAEVKSP
WP_020903468.1 PVLNED--GTPKTV--TQEAKFTREAKVNLVTGEVT---YG-DWS---EAKDLAEVKSP
WP_000287308.1 PVLNED--GTPKTV--TQEAKFTREAKVNLVTGEVT---YG-DWS---EAKDLAEVKSP
WP_033685380.1 PVLSED--GTPKTV--TQEAKFTREAKVNLVSGEVT---YG-DWS---EAKDLPEVKSP
WP_033681547.1 PVLNED--GTPKTV--TQEAKFTREAKVNLVTGEVT---YG-DWT---PAQDLAEVKSP
WP_023948409.1 PVLNED--GTPKTV--TQEAKFTREAKVNLVTGEVT---YG-DWT---PAQDLAEVKSP

WT-GB1-YSIRK -----
WT-GA-YSIRK -----AL-----
YP_005861639.1 LI-----
WP_021874218.1 VV-----
KFL97016.1 VV-----
KFL95230.1 VV-----
WP_033688740.1 TLEKYTAAGITPAVADVPAKTVAATDKDFEETVIYSTKPTTVDPNKPTDPTNPNVTPQ
WP_003097688.1 VV-----
WP_033689127.1 VV-----
WP_020903468.1 VV-----
WP_000287308.1 VV-----
WP_033685380.1 KV-----
WP_033681547.1 VV-----
WP_023948409.1 VV-----

```

Figure A1. Continued.

WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	-----
WP_021874218.1	-----
KFL97016.1	-----
KFL95230.1	-----
WP_033688740.1	PDDVVPNDPKGRITYRELGLIEEVTHTVHYKLEDGSDAGIADNVQTLTFTRTAEVDPVVTGA
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----
WT-GB1-YSIRK	----- EKV -----
WT-GA-YSIRK	----- REF -----
YP_005861639.1	-----NGS-----GTKLTNVVAMVNLFPQASDTSFAFQ
WP_021874218.1	-----SGYHLVSVDRDQDGNNVKDVTLLTHDDNSYIVTVRYA
KFL97016.1	-----EGYVARKTTNNGATVTTVVAGQTKVQQNLEDTVVYD
KFL95230.1	-----EGYVARKTTNNGATVTTVVAGQTKVQQNLEDTVVYD
WP_033688740.1	ISNFGTWKAKGGDTTIDAVTTPNKDGY-----VASAKTSTERTNVAATDKDSEETIIYR
WP_003097688.1	-----KGY-----LAD-KATVPTVNVVTADSKDITEVVITYK
WP_033689127.1	-----TGF-----LAD-KASVPVVNVVTGDSKDITEVVITYK
WP_020903468.1	-----TGF-----LAD-KASVPVVNVVTGDSKDITEVVITYK
WP_000287308.1	-----TGF-----LAD-KASVPVVNVVTGDSKDITEVVITYK
WP_033685380.1	-----DGY-----LAD-KASVAVVNVVTGDSSEDIKEVVITYK
WP_033681547.1	-----KGY-----LAD-KATVPTTKVTADSENTTEVVITYK
WP_023948409.1	-----KGF-----LAD-KASVPVVNVVTGDSSEDIKEVVITYK
WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	LNGRPVYNG-----DKTGTYTFLYTELG-----
WP_021874218.1	KNGKII PVDPNGHPI-----PNVPQPYPTDPNNPAKVTPDEPVPNIPGMTPS----
KFL97016.1	KVGKLV PVGPDGKTP-----IPDAPTPSYPNPDPTDPTKVIPNEPVPDVPGYTPVDP--
KFL95230.1	KVGKLV PVGPDGKTP-----IPDAPTPSYPNPDPTDPTKVIPNEPVPDVPGYTPVDP--
WP_033688740.1	KLGSYVPVVPEG-ITPPADADLNPKYPNATPADPTKPGTPTETPVVPIPGTTPVGPNG
WP_003097688.1	PLGSWVPNIPGQ-PT-----NPIKYPN-DPQDPTKPGQP--TETLPYVPGFTPKDKDG
WP_033689127.1	PLGSWIPNIPGK-TP-----TPIKYPN-NPNDPTKPGD--KPILPYEPGMTPKDGND
WP_020903468.1	PIGSWIPNIPGQ-PT-----NPIKYPN-NPDDPTQPGKP--TEVLPYVPGFTPKDKDG
WP_000287308.1	PIGSWIPNIPGQ-PT-----NPIKYPN-NPDDPTQPGKP--TEVLPYVPGFTPKDKDG
WP_033685380.1	PLGSWVPNIPGQ-PT-----DPIKYPN-DPTDPTTPGTD--KPKVPYVEGFTPKDKDG
WP_033681547.1	PLGSWIPNIPGQ-PT-----DPIKYPN-DPTDPTKPGKD--KPVLPYVPGMTPKDKDG
WP_023948409.1	PLGSWVPNIPGQ-PT-----NPIKYPN-DPTDPTKPGQP--TEVVPYVPGYTPKDGNG

Figure A1. Continued.

WT-GB1-YSIRK	-----EESQTIEGE
WT-GA-YSIRK	-----D-----
YP_005861639.1	--LKSNNPDGTPKDETGYVPADQVTDW-----SKIKSIIIK
WP_021874218.1	--VPTVPTDTPGKDTVPVYTPVAPAKDQVAQVIYRDVNDP-----NKVTQLATS
KFL97016.1	---TPITPEDPTKDTVPVYTKDPVKAGLTVQYIDQ-----DNNNSVIKS
KFL95230.1	---TPITPEDPTKDTVPVYTKDPVKAGLTVQYIDQ-----DNNNSVIKS
WP_033688740.1	KPLTPKDPNDPTKGYEVPDLPTDPTENTTITYVKDGSQVAVTHFIEVNSETDKTEKGAVA
WP_003097688.1	NPLKPVNPNNPEEGYIVPDLPTDPSQDTPINYVKD-TQKAKTTFVD-----EKGNPPIPGV
WP_033689127.1	QPLKPVDPSPDTKGYIVPDLPTDPSQDTPINYVKD-TQKAKTTFVD-----EKGNPPIPGV
WP_020903468.1	NPLKPVDPDPTKGYEVPNLPTDPSQDTPINYVKD-MQKAKTTFVD-----EKGNPPIPGV
WP_000287308.1	NPLKPVDPADPTKGYIVPDLPTDPSQDTPINYVKD-TQKAKTTFVD-----EKGNPPIPGV
WP_033685380.1	NPLKPVDPNDPKEGYEVPNLPTDPSQDTPINYVKD-TQKAKTTFVD-----EKGNPPIPGV
WP_033681547.1	NPLKPVDPNDPTKGYEVPNVPTNPGEDTPINYVKD-TQKAKTTFVD-----EKGNPPIPGV
WP_023948409.1	QPLKPVDPNNPTKGYEVPVPTNPGEDTPINYVKD-TQKAKTTFVD-----EKGNPPIPGV
 WT-GB1-YSIRK	 TSVDS-----
WT-GA-YSIRK	-----
YP_005861639.1	TSSLSNNDRSDRLIFTGIDPNLVNDAGKTGY--ISTGFYSDDTKPFISSLAVYNSSTVAN
WP_021874218.1	GDLTGKAGSEIDYNAQSE----IDNLINKGYVLKNGF--PAGAVFDNDDNKTQTFYIDF
KFL97016.1	DAVDGNIGDKIDYSTASS----ITDFENKGYILVTDGFTGQAGDEFTTENN-GQVYKVV
KFL95230.1	DAVDGNIGDKIDYSTASS----ITDFENKGYILVTDGFTGQAGDEFTTENN-GQVYKVV
WP_033688740.1	ESVVDTGDTGKAFTKAADVTTATIEALKAKGYTVVENNY--PTDGTFDADSKTNQVYKVLV
WP_003097688.1	DAITEEGSDTPTLTKEAEVKAKIKELNKGVELVSNTY--PEGGKFDKDKDQDFEKVTL
WP_033689127.1	AEITEQGSDTPTLTKEAEVKAKIKELNKGVELVSNTY--PEGGKFDKDKDQDFEKVTL
WP_020903468.1	DAITEQGSDTPTLTKEADVAKIKELNKGVELVSNTY--PEGGKFDKDKDQDFEKVTL
WP_000287308.1	DAITEEGSDTPTLTKEADVAKIKELNKGVELVSNTY--PEGGKFDKDKDQDFEKVTL
WP_033685380.1	DAITEEGSDTPTLTKEAEVKAKIKELNKGVELVSNTY--PEGGKFDKDKDQDFEKVTL
WP_033681547.1	DAITEEGSDTPTLTKEAEVKAKIKELNKGVELVSNTY--PEGGKFDKDKDQDFEKVTL
WP_023948409.1	DAITEEGSDTPTLTKEADVAKIKELNKGVELVSNTY--PEGGKFDKDKDQDFEKVTL
 WT-GB1-YSIRK	 -----VYDFNRP
WT-GA-YSIRK	-----
YP_005861639.1	KVKPANITITGEA-NINFKLQYTDENGQL-QTINL-----PDLSTSYNLAQNNT
WP_021874218.1	VHGTVPVPTDTPG-KPGEPIPNPDGPK-WPDGT---SEDSLKKSQTITHYVYSDGSK
KFL97016.1	KHGTRPVTPENPA-DPNEPVDPDHPTPTPSNPNL---SKEDLQKTITRTIEYKYADGTQ
KFL95230.1	KHGTRPVTPENPA-DPNEPVDPDHPTPTPSNPNL---SKEDLQKTITRTIEYKYADGTQ
WP_033688740.1	TAKPITVNPNDPTPTKGQPIDPNNPTGPK-WTPELIKELEDGRTEEVKRTIKYVYADGSK
WP_003097688.1	KAKEVTVTPDQPK-TPGTPVDPNNDGPK-YPAGL---EEKDLNKTVTRTITYVYADGTP
WP_033689127.1	KAKEVTVTPDQPK-TPGTPVDPNNDGPK-YPAGL---EEKDLNKTVTRTITYVYADGTP
WP_020903468.1	KAKEVTVTPDQPK-TPGTPVDPNNDGPK-YPAGL---EEKDLNKTVTRTITYVYEDGTP
WP_000287308.1	KAKEVTVTPDQPK-TPGTPVDPNNDGPK-YPAGL---EEKDLNKTVTRTITYVYADGTP
WP_033685380.1	KERVVPVTPDQPK-TPGTPVDPNNDGPK-YPAGL---EEKDLNKTVTRTITYLYEDGTP
WP_033681547.1	KERVVPVTPDQPK-TPGTPVDPNNDGPK-YPAGL---EEKDLNKTVTRTITYVYEDGTP
WP_023948409.1	KERVVPVTPDQPK-TPGAPVDPNNDGPK-YPAGL---EEKDLNKTVTRTITYVYEDGTP

Figure A1. Continued.

WT-GB1-YSIRK	EIDWSGAEEVYERELQATKEAAIAELQGI-----
WT-GA-YSIRK	-----KYG-----
YP_005861639.1	MLTEQEAIELANKNAASSIPANYEIKSATL-----QSGGKTWQTDAPGTPVFG
WP_021874218.1	AKDDN-----VQSFDFTKSAVVDKVTGEI-----ISQTG-WNVDSHTFGNVDT
KFL97016.1	-----AHELVKQELTFTGKGTIDLVTGNLVTVDEEDGNITSQNGKITWNHESQEFPA
KFL95230.1	-----AHEPVKQELTFTGKGTIDLVTGNLVTVDEEDGNITSQNGKITWNHESQEFPA
WP_033688740.1	-----AADSVQETKEFKRSATINPVTGKV-----TFGD--WS-PAQTFEAVTS
WP_003097688.1	VLNEDGTPKTVTQEAKFTREAKVNLVTGDV-----TYGD--WT-PAQDLAEVKS
WP_033689127.1	VM-ENGAPKVVTQEAKFTREAKVNLVTGEV-----TYGD--WS-EAKDLAEVKS
WP_020903468.1	VLNEDGTPKTVTQEAKFTREAKVNLVTGEV-----TYGD--WS-EAKDLPEVKS
WP_000287308.1	VLNEDGTPKTVTQEAKFTREAKVNLVTGEV-----TYGD--WS-EAKDLPEVKS
WP_033685380.1	VLNEDGTPKVVTQEAKFTREAKVNLVTGEV-----TYGD--WT-PAQDLAEVKS
WP_033681547.1	VLNEDGTPKVVTQEAKFTREAKINLVTGEV-----TYGD--WT-PAQDLAEVKS
WP_023948409.1	VLNEDGTPKTVTQEAKFTREAKVNLVTGEV-----TYGD--WT-PAQDLAEVKS
WT-GB1-YSIRK	-----YI-----KDEELFGRIQDAERIEEVKKLR-----
WT-GA-YSIRK	--VSDYY-----KKLVNSAKTVAGVKKLQAQV-----
YP_005861639.1	GQVQYFY-----NNATVLLQAVPIQRTL-----
WP_021874218.1	PVIDGYH-----ADKRTAGGTTITPDDLNEKVTV-----
KFL97016.1	IDHDGYISSIONSSTASVDGQTGAVGTETVTPNSQNGNIVI-----
KFL95230.1	IDHDGYISSIONSSTASVDGQTGAVGTETVTPNSQNGNIVI-----
WP_033688740.1	PKVTNFT-----PDKESVPAAEVTATAEDINETVIYTTKPANIDPSKPTDP
WP_003097688.1	PVVKGYL-----ADKASVAVVNVTDGSEDIKEVV-----
WP_033689127.1	PVVTGYL-----ADKASIPVVNVTDGSKDITEVV-----
WP_020903468.1	PVVKGYL-----ADKATVPATKVTADSENTKEVV-----
WP_000287308.1	PVVKGYL-----ADKATVPATKVTADSENTKEVV-----
WP_033685380.1	PVVPGYL-----ADKASVPVVNVTDGSKDITEVV-----
WP_033681547.1	PVVKGYL-----ADKATVPPTTKVTADSENTTEVV-----
WP_023948409.1	PVVKGYL-----ADKVTVPPTTKVTADSENTTEVV-----
WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	-----
WP_021874218.1	-----
KFL97016.1	-----
KFL95230.1	-----
WP_033688740.1	NTPNVTPRPDDRVPNDPKGRTYKELGLIEEVTHTVHYKLADGSDAGIPDNVQTLTFRTA
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----

Figure A1. Continued.

WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	-----T
WP_021874218.1	-----T
KFL97016.1	-----TLTRNPDPVPVAAQGSINYIDDTTGQTIESANFSGNVGQKINYTT
KFL95230.1	-----TLTRNPDPVPVAAQGSINYIDDTTGQTIESANFSGNVGQKINYTT
WP_033688740.1	DLDPVTGAISNFGTWTAKDNDTTIDAITTPNKPGYVASAAKSTERTNVQATDKDSEETII
WP_003097688.1	-----T
WP_033689127.1	-----T
WP_020903468.1	-----T
WP_000287308.1	-----T
WP_033685380.1	-----T
WP_033681547.1	-----T
WP_023948409.1	-----T
WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	YQVIDENDP-----
WP_021874218.1	YTPNGKIIP-----
KFL97016.1	AGSIKNWEA-----
KFL95230.1	AGSIKNWEA-----
WP_033688740.1	YRKLGSYVPVPIEGVTPPTGTDLTPKPYENPINEDPTRPGTPTETPVVPYIPGTTPVGPN
WP_003097688.1	YKPLGSWVP-----
WP_033689127.1	YKPIGSWIP-----
WP_020903468.1	YKPIGSWIP-----
WP_000287308.1	YKPIGSWIP-----
WP_033685380.1	YKPLGSWVP-----
WP_033681547.1	YKPLGSWVP-----
WP_023948409.1	YKPLGSWIP-----
WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	-----
WP_021874218.1	-----
KFL97016.1	-----
KFL95230.1	-----
WP_033688740.1	GKPLTPKDPNDPTKGYEVPKVPEDPTQNTTITYVKDGSQVALVHFIAKADGTAVHVSVAEA
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----

Figure A1. Continued.

WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	-----
WP_021874218.1	-----
KFL97016.1	-----
KFL95230.1	-----
WP_033688740.1	GDTGKAIKTTNIDNVKAELEAKGYEVVAPTDAAYTAERVAFYAEANRTFDDKDDKGN DGI
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----
WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	-----
WP_021874218.1	-----VDPN-----
KFL97016.1	-----KGYNLVSNNFKDGEEVFTDGKNAFEVHVLV
KFL95230.1	-----KGYNLVSNNFKDGEEVFTDGKNAFEVHVLV
WP_033688740.1	SQVYYVIVKEGITPIDDPDKPLDPNTPDVT PKPGDKVPGDPKQRTFEQLGLLDEVNRTINY
WP_003097688.1	-----NIPGQP-----
WP_033689127.1	-----NIPGQP-----
WP_020903468.1	-----NIPGQP-----
WP_000287308.1	-----NIPGQP-----
WP_033685380.1	-----NIPGQP-----
WP_033681547.1	-----NIPGQP-----
WP_023948409.1	-----NIPGQP-----
WT-GB1-YSIRK	-----SDAI-
WT-GA-YSIRK	-----V-
YP_005861639.1	-----SNPV-
WP_021874218.1	-----GNPI-
KFL97016.1	HA-----TTPV-
KFL95230.1	HA-----TTPV-
WP_033688740.1	RYANTDKVDADKRGQEARPTVEQKLRYSRKGNLN KVTGEITYTSDWTKPQILAEVTSFVI
WP_003097688.1	-----TSPI-
WP_033689127.1	-----TTPI-
WP_020903468.1	-----TNPI-
WP_000287308.1	-----TNPI-
WP_033685380.1	-----TNPI-
WP_033681547.1	-----TDPI-
WP_023948409.1	-----TTPI-

Figure A1. Continued.

```

WT-GB1-YSIRK -----
WT-GA-YSIRK -----
YP_005861639.1 -----
WP_021874218.1 -----PNVPTPQYPTDPTDP
KFL97016.1 -----TPENPGKPGEP
KFL95230.1 -----TPENPGKPGEP
WP_033688740.1 EGYVADIKAAEKVENVAHDAADSVVNVVYTPLGKYVPKVPEGFEVPKVEKPQYPNDPTDP
WP_003097688.1 -----KYPNDPTDP
WP_033689127.1 -----KYPNDPQDP
WP_020903468.1 -----KYPNDPTDP
WP_000287308.1 -----KYPNDPTDP
WP_033685380.1 -----KYPNDPTDP
WP_033681547.1 -----KYPNDPTDP
WP_023948409.1 -----KYPNDPTDP

WT-GB1-YSIRK -----
WT-GA-YSIRK -----
YP_005861639.1 --TIQPNTDLLNNGQVITGNQGSNV-----
WP_021874218.1 TKV-TPDEPVPNIPGLTPSVPTVT-----
KFL97016.1 VNPTNPDDPHKYPDNYVPQELAKTVTRDVTYVYADGSQAEAPVHQEVKFTGNGYLDLVTG
KFL95230.1 VNPTNPDDPHKYPDNYVPQELAKTVTRDVTYVYADGSQAEAPVHQEVKFTGNGYLDLVTG
WP_033688740.1 TKPGTPTTVIPHVPGTTPKDPNGNPLK-----
WP_003097688.1 TKPGQPTETLPYVPGFTPEDKDG NPLK-----
WP_033689127.1 TKPGQPTETLPYVPGFTPEDKDG NPLK-----
WP_020903468.1 TKPGQPTETLPYVPGFTPEDKDG NPLK-----
WP_000287308.1 TKPGQPTETLPYVPGFTPEDKDG NPLK-----
WP_033685380.1 TKPGSDKPVLPHYVPGMTPKDKDG NPLK-----
WP_033681547.1 TKPGQPTETVPYVPGYTPKDKDG NPLK-----
WP_023948409.1 TKPGKPTDVLPHYV-----

WT-GB1-YSIRK -----
WT-GA-YSIRK -----
YP_005861639.1 -----PNDATTAYDAVKNALE-----
WP_021874218.1 -----PTDPGKDTVPYNPVPAKDQAADVNYV-----DADEDNKLITSSGDLTG
KFL97016.1 EYVTVDNNGKITGKGQINWTPESANFDATKSIDTSKYQIVGIKENNTANVDQTTGVVAG
KFL95230.1 EYVTVDNNGKITGKGQINWTPESANFDATKSIDTSKYQIVGIKENNTANVDQTTGVVAG
WP_033688740.1 ---PVDPNDPKSGY-VPPTPENPTEDTQI--TYEKDTQKAKVTYVVEGTGTVLHTDNLEG
WP_003097688.1 ---PVDPT-----
WP_033689127.1 ---PVDPTDPSKGYVVPNIPTDPSQDTVI--NYVANKANLVVKYVDENGKDLIPSETTEG
WP_020903468.1 ---PVDPNDPKGYEVPSIPTNPGEDTPI--NYVANKANLVVKYVDENGKELQPTETKEG
WP_000287308.1 ---PVDPNDPKGYEVPSIPTNPGEDTLI--NYVANKANLVVKYVDENGKELLPTETKEG
WP_033685380.1 ---PVEPNDPKGYEVPSVPTNPGEDTPI--NYVKDKQAKTTFVDEKGNPIPGVDAITE
WP_033681547.1 ---PVDPNDPKGYEVPNVPTNPGEDTPI--NYVANKANLVVKYVDEKGDLLPAETTEG
WP_023948409.1 -----

```

Figure A1. Continued.

```

WT-GB1-YSIRK -----
WT-GA-YSIRK -----
YP_005861639.1 -----
WP_021874218.1 KAGETINYSTADTI-----KDLENKGYVLVNDGF--PA-----
KFL97016.1 ETVTQNSNNSSVVITLANKPAPVVEKGSITVKVHDLTDNVDLPQYGKESGEQEVGTSFTY
KFL95230.1 ETVTQNSNNSSVVITLANKPAPVVEKGSITVKVHDLTDNVDLPQYGKESGEQEVGTSFTY
WP_033688740.1 KSGEPIEYSTVTKL-----AELKALGYDLVTDGFTTAT-----
WP_003097688.1 -----
WP_033689127.1 KVGD--EYTTTGKV-----IPGHLLVRVEG--ES-----
WP_020903468.1 KVGD--DYSTSGKV-----ITGYVLDRVEG--EA-----
WP_000287308.1 KVGD--DYSTSGKV-----ITGYVLDRVEG--EA-----
WP_033685380.1 EGDS--DTPLTKEAEVKAKIKELENGKGYELVSNTY--PE-----
WP_033681547.1 KVGD--EYATSGKV-----IKGYVLVRVDG--EA-----
WP_023948409.1 -----

WT-GB1-YSIRK -----
WT-GA-YSIRK -----
YP_005861639.1 -----
WP_021874218.1 -----GAKYDSDDNNTTQIY-----TVVLKHGTTTTITPDKPGKPGEPINPN
KFL97016.1 DKNVITELINKGYKLVDGGENVPSEVAKGAKTITILVEHDTVPTPENPGKPGEPINPN
KFL95230.1 DKNVITELINKGYKLVDGGENVPSEVAKGAKTITILVEHDTVPTPENPGKPGEPINPN
WP_033688740.1 -----DKNYDKDKTKVDQSFVTVKPHVEPIKPVDPENPNPNPGDPIDPN
WP_003097688.1 -----
WP_033689127.1 -----KGKIGKDGS-----
WP_020903468.1 -----KGKIGTDGTTVTVYVKPLGSWIPNIPGQPTNPIKYPN--DPTDPTKPG
WP_000287308.1 -----KGKIGTDGT-----
WP_033685380.1 -----GGKFDKDKDQEFKVT LKERVVPVTPDQPK-----TPGAPVDPN
WP_033681547.1 -----KGKIGKDGS-----
WP_023948409.1 -----

WT-GB1-YSIRK -----
WT-GA-YSIRK -----
YP_005861639.1 -----
WP_021874218.1 DPDGPKWPDNSGEN-NLSKTGTQTIHYT-----GAGDKTPEDNKQEFFTTKTM
KFL97016.1 DPDGPKWPE-GTDENSVKRTGTQTIHYE-----GAGDKTPSDDVQTFDFTKKM
KFL95230.1 DPDGPKWPE-GTDENSVKRTGTQTIHYE-----GAGDKTPSDDVQAFDFTKKM
WP_033688740.1 NPDGPKWTEDLIKKIDTTRHVNRTITYVNEKGEEV-----AKKVTDKVFTFTREGKINTV
WP_003097688.1 -----
WP_033689127.1 -----TVTYV-----
WP_020903468.1 QPTETLPYVPGYTPKDGNGQPLKPVDQDPTKGYVVPNIPTDPSQDTVINYEANKAKLVV
WP_000287308.1 -----TVTYV-----
WP_033685380.1 NPDGPKYPAGLEEK-DLNKTVIRITITYVYEDGTPVLNEDGTPKVVTQEAFTREAKVNLV
WP_033681547.1 -----TVTYV-----
WP_023948409.1 -----

```

Figure A1. Continued.

```

WT-GB1-YSIRK -----
WT-GA-YSIRK -----
YP_005861639.1 -----
WP_021874218.1 VVDNVTGKVITDGAWNVTSHTFGNVDTFVIDGYHADKRTAGGTTITPDDLNKTVTVNYTP
KFL97016.1 LVDKVTGKIIDSGEWNVTSHTFGYKDTFVIDGYHADKRNAGGSVVTFNDLNKKVVVYKYP
KFL95230.1 LVDKVTGKIIDSGEWNVTSHTFGYKDTFVIDGYHADKRNAGGSVVTFNDLNKKVVVYKYP
WP_033688740.1 TGEITYGDWTAkdGDTTFDKVESPVVKGYILKDAKQKEVAATTGLTADSKDENIKVVVYP
WP_003097688.1 -----
WP_033689127.1 -----YKP
WP_020903468.1 KYVDENGKDLIPSETTEGKVGDEYTTSGKVI PGHLLVRVDGDAKGKIGTEGSTVTVYVYKYP
WP_000287308.1 -----YKP
WP_033685380.1 TGEVTVGDWTPAQD-----LAEVKSPVVKGYLADKVTVPPTTKVTADSKDTKEVVYKYP
WP_033681547.1 -----YKP
WP_023948409.1 -----

WT-GB1-YSIRK -----
WT-GA-YSIRK -----
YP_005861639.1 -----
WP_021874218.1 NGKIIPVDPNG-----NPIPNVPTP-----QYPTDPTDPTKVTPD--EPVPTIPG
KFL97016.1 NGKIIPVDPNG-----NPIPNVPTP-----TYPTDPTDPTKVVPD--EPVPDIPG
KFL95230.1 NGKIIPVDPNG-----NPIPNVPTP-----TYPTDPTDPTKVVPD--EPVPDIPG
WP_033688740.1 VGKVTITVPPGVTPPTFVTDTPYENVPGEKGVVPPSPPTKPKQDPNSPKVPVPIPHIPG
WP_003097688.1 -----
WP_033689127.1 IG-----SWIPNIPGQPTTPI-KYPNDPQDPTKPGQPT-EVLPHYVPG
WP_020903468.1 IG-----SWIPNIPGQPTNPI-KYPNDPQDPTKPGQPT-EVLPHYVPG
WP_000287308.1 LG-----SWIPNIPGQPTNPI-KYPNDPQDPTKPGQPT-ETLPHYVPG
WP_033685380.1 LG-----SWIPNIPGQPTNPI-KYPNDPADPTKPGSDK-PVLPHYVPG
WP_033681547.1 LG-----SWVPNIPGQPTNPI-KYPNDPQDPTKPGQPT-EVVPYVPG
WP_023948409.1 -----

WT-GB1-YSIRK -----
WT-GA-YSIRK -----
YP_005861639.1 -----AKGYVISSKSSTVPTIFGP-----
WP_021874218.1 YTPS-----TPTVTPDTPGKDTVPVYNPVVPAKN
KFL97016.1 MTPS-----TPTVTPEDPGKDTVPVYNPVVPAKN
KFL95230.1 MTPS-----TPTVTPEDPGKDTVPVYNPVVPAKN
WP_033688740.1 TTPVVPKDPKPKISPDPNPLVPLTPVDPNDPTKGYEVPPV-PTDPSTDPITY---VTDK
WP_003097688.1 -----DPTKGYVVPNI-PTDPSQDTVINY---VANK
WP_033689127.1 FTPE-----DKDGNPLKPVDPKDPKGYVVPNI-PTDPSQDTVINY---VANK
WP_020903468.1 FTPE-----DKDGNPLKPVDPDPTKGYVVPNI-PTDPSQDTVINY---VANK
WP_000287308.1 FTPE-----DKDGNPLKPVDPNDPTKGYEVPSI-PTNPGEDTPINY---VANK
WP_033685380.1 HTPV-----DGNGQPLKPVDPKDPKGYEVPNV-PTNPGEDTPINY---VANK
WP_033681547.1 YTPK-----DKDGNPLKPVDPNDPTKGYEVPSV-PTNPGEDTPINY---VANK
WP_023948409.1 -----

```

Figure A1. Continued.

WT-GB1-YSIRK	----- DARQ ----- ALVGKLLDDQEASA -----
WT-GA-YSIRK	----- ESAK -----
YP_005861639.1	-----DNTALTIYVTHKTIINVSTPDQWPST-----
WP_021874218.1	QKAVVNYVDAEDN-KLITSSGDLTGKAGKKIDYS-----
KFL97016.1	QVAQVIYRDVQDGANKQLATSGDLTGKSGSEISYS-----
KFL95230.1	QVAQVIYRDVQDGANKQLATSGDLTGKSGSEISYS-----
WP_033688740.1	QKAITNFV--DEKG-KVVSTPVVDEGDSGANFTKSKVDEVTKTIEKLEKAGYRVVKNEFP
WP_003097688.1	AKLVVKYV--DENG-KDLIPAETTEGKVGDEYTTT-----
WP_033689127.1	AKLVVKYV--DENG-KDLIPSETTEGKVGDEYTTT-----
WP_020903468.1	AKLVVKYV--DENG-KDLIPAETTEGKVGDEYTTT-----
WP_000287308.1	ANLVVKYV--DENG-KELLPTETEGKVGDDYSTS-----
WP_033685380.1	ANLVVKYV--DENG-KELLPSETTEGKVGDEYATS-----
WP_033681547.1	ANLVVKYV--DEKG-KDLLPAETTEGKVGDEYATS-----
WP_023948409.1	-----
WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	-----
WP_021874218.1	-----TSSTIEDLINKGYVLVNDGFPKDATYDNDNTTQTYTVV
KFL97016.1	-----TADQIKKLINQGYVLKNDGFP
KFL95230.1	-----TADQIKKLINQGYVLKNDGFP
WP_033688740.1	SKDTRVFDKDKSVDQIFTVTVAERII PVT PGKPVDPNDPNLPKNPDGTPVTPSTPE---
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----
WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	-----
WP_021874218.1	FKHGTVPVTPTPNPGKPGEPIPNPDGPKWPDGTGENSIDKTVTRTITFVDSNGKEVSSP
KFL97016.1	AGAVFDNDDSKNQVFYVDFIHGQAPVNPDPNPHGIDPSQYEKTVTEKVHYVGAGDKTPAD
KFL95230.1	AGAVFDNDDSKNQVFYVDFIHGQAPVNPDPNPHGIDPSQYEKTVTEKVHYVGAGDKTPAD
WP_033688740.1	-----PGKPVFPDGNPSPVWPNTVKDLVTEKSATRTIKYVDRNGKEVSET
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----

Figure A1. Continued.

```

WT-GB1-YSIRK -----
WT-GA-YSIRK -----
YP_005861639.1 -----STDADKVS
WP_021874218.1 VEQSVHFTATGVIDKVTGKWVTPLSWSPDQSIDGRNVFVVDGYHVVSIDK--DGNGLTNV
KFL97016.1 NVQNSKWTRTLTIDTVTGKVVENGQYTTDWSIAKGEKTVYDQVSTPVIDGYHADKREVPA
KFL95230.1 NVQNSKWTRTLTIDTVTGKVVENGQYTTDWSIAKGEKTVYDQVSTPVIDGYHADKREVPA
WP_033688740.1 RTETIKFTREAKVNIVTGEITYGEWT-----TDRNDDIFNGYQVPVVKGYIAKAGDLES
WP_003097688.1 -----GKVIPGYVLVRVDG--EAKGKIGT
WP_033689127.1 -----GKVIPGYLLVRVEG--EAKGKIGK
WP_020903468.1 -----GKVIPGYLLVRVDG--DAKGKIGK
WP_000287308.1 -----GKVITGYVLDRVEG--EAKGKIGT
WP_033685380.1 -----GKVITGYVLERVEG--EAKGKIGE
WP_033681547.1 -----GKVIKGYVLVRVDG--EAKGKIGK
WP_023948409.1 -----

WT-GB1-YSIRK -----DSEYDFVRPDVD--
WT-GA-YSIRK -----
YP_005861639.1 LSKTI-----TRTITVEGLPTA-----VEGTTQTVTFTRTAVVDE
WP_021874218.1 KGVTL---THDDSSYKVTVTYVQNGKIIPVDPNGKPIPNVPTPQYPTDPTDPSKVVPNEP
KFL97016.1 TAVTQ-----DDIEVTVTYKPNGKIIPDPSNPPIPNVNPNTYPTDPTDPTKVVDPQF
KFL95230.1 TAVTQ-----DDIEVTVTYKPNGKIIPDPSNPPIPNVNPNTYPTDPTDPTKVVDPDEP
WP_033688740.1 STKDVQVTPDTIKDINETVIYDKLGSWIPNIPGTPTNP----ITYPNDPKDPTKPGTDKP
WP_003097688.1 DGSTV-----TYVYTPLGSWVPNIPGQPTSP----IKYPNDPTDPTKPGSDKP
WP_033689127.1 DGSTV-----TYVYKPIGSWIPNIPGQPTNP----IKYPNDPADPTKPGSDKP
WP_020903468.1 EGSIV-----TYVYKPIGSWIPNIPGQPTNP----IKYPNDPQDPTKPGSDKP
WP_000287308.1 DGTTV-----TYVYKPLGSWIPNIPGQPTTP----IKYPNDPQDPTKPGQPTPE
WP_033685380.1 NGTTV-----TYVYKPLGSWVPNIPGQPTDP----IKYPNDPTDPTKPGNDKP
WP_033681547.1 DGSTV-----TYVYKPLGSWVPNIPGQPTDP----IKYPNDPTDPTKPGKDKP
WP_023948409.1 -----

WT-GB1-YSIRK -----
WT-GA-YSIRK -----
YP_005861639.1 VTGKVIGYV-----
WP_021874218.1 V-PTIPGYTPSTPTVTPTDPGKDTVPYPNPVEAKQGSVQVIFHDDTTNTTIPDVGYNSGS
KFL97016.1 V-PEVPGMT-----
KFL95230.1 V-PNIPGMTPTSTPTVTPEDPGKDTVPYPNPVPAKDQAAVVNYVDADNNNTIITSSGNLT
WP_033688740.1 KVPYVPGFI-----
WP_003097688.1 VLPYVPGYT-----
WP_033689127.1 VLPYVPGHT-----
WP_020903468.1 VLPYVPGHT-----
WP_000287308.1 VLPYVPGFT-----
WP_033685380.1 VLPYVPGYT-----
WP_033681547.1 VLPYVPGYT-----
WP_023948409.1 -----

```

Figure A1. Continued.

WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	-----
WP_021874218.1	QKAGTKVDYTTTKSISDLEVKGYYVSTDGTIPTEITADKNIT--VTVHMKHGTTTVPD
KFL97016.1	-----
KFL95230.1	GKAGSRIDYSTKTTIADLENKGYVLVNDGFADATFDNDDSTTQVFTVVLKHGTVPVTP
WP_033688740.1	-----
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----
WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	-----
WP_021874218.1	KPGKPGEPINPNPDGPKWPDTTGKDNLSKTGTQTIHYTGAGNNTPKDNVQSFTFTRTAV
KFL97016.1	-----
KFL95230.1	NPGKPGEPINSNDPDGPKWPEGTDENSVKRTGTQTIHYVGAGDKTPSDDVQTFDFTRKMV
WP_033688740.1	-----
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----
WT-GB1-YSIRK	-----
WT-GA-YSIRK	-----
YP_005861639.1	-----
WP_021874218.1	VDNVTGKVIISTGAWNVTSHTFGNVNTPVVDGYHADKRTAGNTTITPEDLNKTVTVNYTAN
KFL97016.1	-----
KFL95230.1	VDKVTGKVVDGGSWNVTSHTFGYKNTPVIDGYHADKRNAGGSVVTPTDDLNTVTVTYKQN
WP_033688740.1	-----
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----

Figure A1. Continued.

```

WT-GB1-YSIRK -----
WT-GA-YSIRK -----
YP_005861639.1 -----DPSDTSQ-----
WP_021874218.1 GKIIIPVDPNGKPIPNVPTPTYPTDNDPTKVVVNEPVPTIPGYKPSVPTVTPSDPGKDTF
KFL97016.1 -----PSTPTV-----TPEDPGKDTF
KFL95230.1 GKIVPVDPSGNPIPNVNPPTYPTDPTDPTKVVVDQFVPEVPGMTPSTPTVTPEDPGKDTF
WP_033688740.1 ----PVDPEGQPLK-----PVDNDPTK---GYEVPD-----VPGDPTQDTP
WP_003097688.1 ----PVDGNGQPLK-----PVDNDPTQ---GYEVPN-----VPNDPTKDTF
WP_033689127.1 ----PVDGNGQPLK-----PVDNDPTK---GYEVPD-----IPTNPGEDTF
WP_020903468.1 ----PVDGNGQPLK-----PVDNDPTK---GYISPD-----IPTNPGEDTF
WP_000287308.1 ----PEDKDGNNPLK-----PVDPKDPSK---GYVVPN-----IPTNPGEDTF
WP_033685380.1 ----PKDKDGNNPLK-----PVDNDPTK---GYEVPN-----VPTNPSEDTF
WP_033681547.1 ----PKDKDGNNPLK-----PVDNDPTK---GYEVPS-----VPTNPGEDTF
WP_023948409.1 -----

WT-GB1-YSIRK -----
WT-GA-YSIRK -----
YP_005861639.1 -----
WP_021874218.1 VPY-----
KFL97016.1 VPYNPVKNPDKVTTVGKQIVHFVDGNGNTPLRDPNTQTHEFKITNGVPDESSHTFTLV
KFL95230.1 VPYNPVKNPDKVTTVGKQIVHFVDGNGNTPLRDPNTQTHEFKITNGVPDESSHTFTLV
WP_033688740.1 INY-----
WP_003097688.1 INY-----
WP_033689127.1 INY-----
WP_020903468.1 INY-----
WP_000287308.1 INY-----
WP_033685380.1 INY-----
WP_033681547.1 INY-----
WP_023948409.1 -----

WT-GB1-YSIRK -----
WT-GA-YSIRK -----
YP_005861639.1 -----
WP_021874218.1 ----APQTTPTVTPNIPVTPNEPSTPTTPTDTSAPTTPHGEDVPVTPNEPDTPA-----
KFL97016.1 DVPVIPGYVAEVKSAGGKTVPDTPPLAEVTVVYHKVGKIVPVDPNGNPIPNVPTPSYTND
KFL95230.1 DVPVIPGYVAEVKSAGGKTVPDTPPLAEVTVVYHKVGKIVPVDPNGNPIPNVPTPSYTND
WP_033688740.1 ----IPK-----DPTPNPTPYP-----GPTPAPTPKPEPKPE-----
WP_003097688.1 ----VP-----APQNPPTP-----APTPEPKPEPKPE-----
WP_033689127.1 ----IPN-----SPKPNPTPYP-----GPTPAPTPKPEPKPE-----
WP_020903468.1 ----IPNVTPNGDQNGYTPQPKPQPEQVVVYYYVDENGKDIAPSEKGAQAPK----GISGY
WP_000287308.1 ----IPNVTPNGDQDGYTPQPKPQPEQVVVYYYVDENGKDIAPSEKGAQAPK----GISGY
WP_033685380.1 ----IPN-----SPKPNPTPYP-----GPTPAPTPKPEPKPE-----
WP_033681547.1 ----VPN-----
WP_023948409.1 -----

```

Figure A1. Continued.


```

WT-GB1-YSIRK -----
WT-GA-YSIRK -----
YP_005861639.1 -----
WP_021874218.1 -----
KFL97016.1 -----PAPHGEKPEEPDRPA-----PAPHAPKAP-----
KFL95230.1 PTDPTKVVPNEPVPVAITGKTPDKTSVTPVDPTKDTPVVYKNNEVPATPNSQKAVVNFIDV
WP_033688740.1 PTDPTKVVPNEPVPVAITGKTPDKTSVTPVDPTKDTPVVYKNNEVPATPNSQKAVVNFIDV
WP_003097688.1 -----PKPEPKPEPETPQPVTADDGDNN---GNNNGTPSTPAQP-----
WP_033689127.1 -----TPQPVTPADNGDNN---GNNNETPTTPAQP-----
WP_020903468.1 -----PAPVPSTPETPEQPVAPVQPEQPTTPTQP-----
WP_000287308.1 EYVTTTKDPNGNLVHHYKKVATPQVPSTPETPEQPVAPVQPEQQTNPQNQP-----
WP_033685380.1 EYVTTTKDPNGNLVHHYKKVATPQVPSTPETPEQPVAPVQPEQPTTPTQP-----
WP_033681547.1 -----PAPVPSTPETPEQPVAPVQPEQPTTNPQNQP-----
WP_023948409.1 -----PREVEKPAKPAQP-----

WT-GB1-YSIRK -----
WT-GA-YSIRK -----
YP_005861639.1 -----
WP_021874218.1 -----
KFL97016.1 NTGKLIKTSGILSGRPGEDINKLYSSAEVIKQLEEAGYEVVYNADFDDGVDGVTKYFDDDDNT
KFL95230.1 NTGKLIKTSGILSGRPGEDINKLYSSAEVIKQLEEAGYEVVYNADFDDGVDGVTKYFDDDDNT
WP_033688740.1 -----
WP_003097688.1 -----
WP_033689127.1 -----
WP_020903468.1 -----
WP_000287308.1 -----
WP_033685380.1 -----
WP_033681547.1 -----
WP_023948409.1 -----

WT-GB1-YSIRK -----
WT-GA-YSIRK -----
YP_005861639.1 -----
WP_021874218.1 -----TAKGNNTPVKENKTVPATAA-PVVK
KFL97016.1 TQQFTVALKLKEKAKTPYPVVPAPETPAKEPEAPAEKVSREQPVKQNVSVPTPQKPVEK
KFL95230.1 TQQFTVALKLKEKAKTPYPVVPAPETPAKEPEAPAEKVSREQPVKQNVSVPTPQKPVEK
WP_033688740.1 -----AAPS-----TP
WP_003097688.1 -----AAPS-----TP
WP_033689127.1 -----AVPT---PAETSVATDSATQTATP
WP_020903468.1 -----AVPA---PAETSVATDSATQPATP
WP_000287308.1 -----AVPT---PAETSVPTDSATKPATP
WP_033685380.1 -----AVPA---PAETSVATDSATQPATP
WP_033681547.1 -----SKQETP
WP_023948409.1 -----

```

Figure A1. Continued.

```

WT-GB1-YSIRK      -----LTGLDEGYARE-----
WT-GA-YSIRK      -----
YP_005861639.1    -----TITDGDNAWTSVNNTWSAFTP-----
WP_021874218.1    NEQTPEAELPQTGEKNDSSAAAILGATAGMIGLIGLLGVKKKHSEN-
KFL97016.1        KTNNKKEVLPQTGADNNEAASILGAVATAIGMTSLIGAKRRKKDDK
KFL95230.1        KTNNKKEVLPQTGADNNEAASILGAVATAIGMTSLIGAKRRKKDDK
WP_033688740.1    QYMDGQRELPNTGTEDHASLAALG-LLGALSGFGLI-ARKKREDEE
WP_003097688.1    QYMDGQRELPNTGTEDNASLAALG-LLGVLSGFGLV-ARKKKED--
WP_033689127.1    KYVDGQKELPNTGTEANASLAALG-LLGALGGFGLL-TRKKKED--
WP_020903468.1    KYVDGQKELPNTGTEANASLAALG-LLGALGGFGLL-ARKKKED--
WP_000287308.1    KYVDGQKELPNTGTEANASLAALG-LLGALGGFGLL-ARKKKED--
WP_033685380.1    KYVDGQKELPNTGTEANASLAALG-LLGALGGFGLL-SRKKKED--
WP_033681547.1    KYVEGQKELPNTGTEANASLASLG-LLGALGGIGLL-TRKKKED--
WP_023948409.1    -----

```

Figure A1. Continued.

```

WT-GA
WT-GB1
YP_005861639.1 -----MGMFFNKKDNDKQRFGRKLTIGACSVLLSTLIL-----GI
WP_033688740.1 -----MYKSKRRKNKSFWDYWG-LSQRFSIRKYHFGAASVLLGTALILGAAQTTA
WP_003097688.1 -----MEKFHGRKAQRFSIRKYSFGAASVLLGTALFLGAN--GV
WP_033689127.1 -----MYSRMEKYHGRRRAQRFSIRKYSFGAASVLLGTALLGAN--AV
WP_020903468.1 -----MYSRMEKYHGRRRAQRFSIRKYSFGAASVLLGTALVLGAN--GV
WP_000287308.1 -----MYSRMEKYHGRRRAQRFSIRKYSFGAASVLLGTALLGAN--AV
WP_033685380.1 -----MYSRMEKYHGRRRAQRFSIRKYSFGAASVLLGTALVLGAN--GV
WP_033681547.1 -----MYSRMEKYHGRRRAQRFSIRKYSFGAASVLLGTALLGAN--AV
WP_023948409.1 -----MYSRMDKYHGQKVQRFSIRKYSFGAASVLLGTALLGTN--AV
WP_021874218.1 -----MVSKNNRDKKMEAVAERKPHFAIRKLTIGAASVLLGTSLWMSTSTSTV
KFL97016.1 MIYYDSKWRETGMLSKNNYQERLRKMDKQERFSIRKFSVGAASVLVGTAILSMQNVQTV
KFL95230.1 MIYYDSKWRETGMLSKNNYQERLRKMDKQERFSIRKFSVGAASVLVGTAILSMQNVQTV

WT-GA
WT-GB1
YP_005861639.1 GTQEONSKAQAATTETSNTASS-----TDLVDNHRNKNTYLSSEVNET
WP_033688740.1 KAEETVT--ENKTEAVASAPKDDKASENVNTVTPALSATTEAAVVEKPTLSDEEVAKLA
WP_003097688.1 RADEATPSVNPATSGLSNSDKNVSGSTLSTPVVEKLPELKIDAVKADENAEVKEESKNEV
WP_033689127.1 KADETLP-VNPTASDLAATNKKDADSALTTPVVEELPELKIDAVKADEKAKAKEDAKTEA
WP_020903468.1 QAEETVA-VNPATSELSNSDKNLGSGSTLSTPVVEELPELKIDDVKAEEKTEAKEDAKTAA
WP_000287308.1 KADETST-ASTKTSEVTNSDKQKPDSAITTPVVEELPELKIDAVKADEKPEVKEEAK---
WP_033685380.1 QADETLP-VNPATSDLAATNKKDADSALTTPVVEELPELKIDAVKADEKAEPKEDVKTEA
WP_033681547.1 KADETST-ASAKTPEVTNSDKQKSDSGSTTPVVEELPELKIDAVKADEKPEVKEEAKTEA
WP_023948409.1 KADETNT-GSAKATEITNPDQKQPDSAITTPVVEELPELKIDAVKADEKSEVKEEAK---
WP_021874218.1 HADETDNNSDAKTNLESNQSASTGH-----VEKVVEQNQTANENTDDSTKT
KFL97016.1 HADATTDTEKGTDDVTSKNDEQNKKQK-----AYNQVVEDQNKASKTTDTTMV
KFL95230.1 RADATTDTEKGTDDVTSKNDEQNKKQK-----AYNQVVEDQNKASKTTDTTMV

WT-GA
WT-GB1
YP_005861639.1 ATKVNEKETSAGNEQQDSQSAVAQDKQSGEKAADVVTNNRSTVED-KSSNNAESSETTDN
WP_033688740.1 AEASKKDDKASETATTEKTEAADKEKATLTAPLTDKKADKAVDEKADKKDEKKAENPITA
WP_003097688.1 TTVAEKEVTEATTDKTDKKTETDVKEKSDKEHADKTEADKEKT-----EKVETEKAQDD
WP_033689127.1 TPVVEKEITEATAEKTDKKLETDVKEKSDKEQADKKEATKEKTDK-KTSEKVETEKVQDD
WP_020903468.1 TPVAEKEVTEATADKTDKKLETDVKEKSDKEQSDKKEADKEKTNK-ETSEKVETEKAQDD
WP_000287308.1 -PVAEKEVTDKAA-----TEKSDKEQADKKEVAKEKTDK-ESPKKAATEKAQDE
WP_033685380.1 TPVVEKEVSDKSDKEAS-----KEKSDKEQADKKEATKEKTDK-ETSEKVETEKAQDD
WP_033681547.1 KPVAEKEVADKAA-----TEKSDKEQADKKEADKEKTNK-ETSEKATTEKAQDE
WP_023948409.1 -PVAEKEVTDKAA-----TEKSDKEQADKKEVAKEKTDK-ETSEKAETVKPKDE
WP_021874218.1 NNVSAQNTQESVDESSDISSDNAQQNKAITSEEQNSDAAVTIDNN-QAADENKAETQKVT
KFL97016.1 GQDSKVASFSASKNEGTF-----AEASSEDKTSSTTDATQN-KASENTKATEDNKV
KFL95230.1 GQDSKVASFSASKNEGTF-----AEASSEDKTSSTTDATQN-KASENTKATEDNKV

```

Figure A2. MUSCLE alignment of 12 mucus-binding proteins with WT-GB1 and GA. Accession numbers correspond to mucus-binding proteins found in Figure 40.

WT-GA	-----MEAVDANSIAQAKEAA
WT-GB1	-----
YP_005861639.1	TKNTVN-----SDRAEVTNKEANTQTDSKKVTQKTSQAVNDVNKNVAETTTDTR
WP_033688740.1	TKTVLEQLTSEAEVLNTTASNFDKKAEDKAGKEAIATAVASAKVQIEASKKALAAGEIT
WP_003097688.1	VKTVLTQLTSEADVMATVASNFSDEKVTGVEDKQNLAAITAVKLEATASKELL-LSDAS
WP_033689127.1	VKTVLTQLTSEADVMASVASNFSDEKVKDDASKQKLSAAVAAVKLEAVASKGLL-SSDAS
WP_020903468.1	VKTVLTQLTSEADVMATVASNFSDEKVKDVESKQKLSAAIAAVKLEAVASKGLL-SSDAS
WP_000287308.1	VKTVLTQLTSEAEVMASVASNFSDEKVKDEAAKKELAVTIEAVKLEAAKSNDLL-SSDAS
WP_033685380.1	VKTVLTQLTSEAEVMATVASNFSDEKVKDEAKQKLSAAIAAVKLEAVAAKGLLYSNDST
WP_033681547.1	VKTVLTQLTSEAEVMAKVASNFSDEKVKDEAAKKELAVKIEAVKLEAAKSNDLL-SSDAS
WP_023948409.1	VKTVLTQLTSEAEVMATVASNFSDEKVKDEAKQKLSAAIAAVKLEAVASKGLL-SSDAS
WP_021874218.1	DKTTKTKQDDNKSSQITDNKKSSEKAATDTSNKNVVEQSANSVENNANIDNSIAANTQTD
KFL97016.1	DAVADKSTEDKNTATTQESSDNKSTENKTTDAQKVESKVATAKATTTDANSVKTASTTS
KFL95230.1	DAVADKSTEDKNTATTQESSDNKSTENKTTDAQKVESKVATAKATTTDANSVKTASTTS
 WT-GA	 IKEL -----
WT-GB1	-----
YP_005861639.1	NTKVSSYTSNL--DLENIQESLEQ-----QAKENNGKA-
WP_033688740.1	KQELDAQLQRISAAIEAVYDEMCR-----AGHLGKVEAV
WP_003097688.1	KDQMVAQVNRLSAAIEAVYAEMKR-----AGHAGKVEAV
WP_033689127.1	KDQMVAQVNRLSAAIEAVYAEMKR-----AGHAGKVESV
WP_020903468.1	KDQMVAQVNRLSAAIEAVYAEMKR-----AGHAGKVEAV
WP_000287308.1	KDQMVAQVNRLSAAIEAVYTEMKR-----AGHAGKVESV
WP_033685380.1	EEQLTSQVNRISAAIEAVYAEMKR-----AGHAGKVEAV
WP_033681547.1	KDQMVAQVNRLSAAIEAVYAEMKR-----AGHAGKVEAR
WP_023948409.1	KDQMVAQVNRLSAAIEAVYAEMKR-----AGHAGKVEAV
WP_021874218.1	ITKSNIQNLNLSLPSIAQAGQNGKTIKVDNDTTTQELKIGDLSSDLSGDALKANLTGKNQV
KFL97016.1	TDQTTSVNTTTTFN-----
KFL95230.1	TDQTTSVNTTTTFN-----
 WT-GA	 -----
WT-GB1	-----
YP_005861639.1	-----LDAKSVTSLRSDAANFQVLAALTENIAE-----
WP_033688740.1	LAGETTTNTAIVAPTTKTPVKNINKLTDEEIAAVKREIMNANPSITDPSMIEVVQDNGT
WP_003097688.1	IDDATTDQDKIVGKDVVKDGQKVKSVTNGYITMNADNTAP-----
WP_033689127.1	LADTAS---KITGKDILRDGETVNAVTVNAYVDMNADNTAP-----
WP_020903468.1	LADTAS---KITGKDVLDGETVNAVTVNAYVEMNADNTRP-----
WP_000287308.1	LAATAS---KITGKDVLDGETVNAVTVNAYVDMNADNTKP-----
WP_033685380.1	LGATDAKDQITITGKGVIRNGNAIVTVNNAYVTMNADNTAP-----
WP_033681547.1	LAETHT-----GKALIKEGKAVVNVQNAVITMNADNSAP-----
WP_023948409.1	LAETHT-----GKALIKEGKAVVNVQNAVITMNADNSAP-----
WP_021874218.1	LLNQSNSSSEVVVGKNVDPTKQLQAMARTAMFAAVNPNAADNYTTVSDFNALQQAVNDYSV
KFL97016.1	--NQSSTAVLFSASALSSESKALAATPRASQATTNAQAKNNNYKLVTSSELQQAINS-GV
KFL95230.1	-TNQSSTAALFSASALSSESKALAATPRASQATTNAQAKNNNYKLVTSSELQQAINS-GV

Figure A2. Continued.

```

WT-GA
WT-GB1
YP_005861639.1
WP_033688740.1 AG---GATVTINGVKTNIIPSGDTPVGTAGTKNLEQLKNNINWFDFAAASITYSNGTVVGP
WP_003097688.1 -----KAWGFDSTFDTSDLQAG-----
WP_033689127.1 -----TGWGFDTTISTSTLNAG-----
WP_020903468.1 -----TGWGFDTTISTSTLKAG-----
WP_000287308.1 -----VGWGFDTTISTSTLKAG-----
WP_033685380.1 -----KAWGIDVVFDTSQAQNG-----
WP_033681547.1 -----TAWGIEISFDTSRTQKG-----
WP_023948409.1 -----TAWGIEISFDTSRTQKG-----
WP_021874218.1 SGVNISGNITAYG-DLNINRTFTIKGADNNATLSLGQNKIN-----
KFL97016.1 AGINIDRSIDASNVNLAITNTFAIVGINDAAVNLGQKSLN-----
KFL95230.1 AGINIDRSIDASNVNLAITNTFAIVGINDAAVNLGQKSLN-----

WT-GA
WT-GB1
YP_005861639.1
WP_033688740.1 ARKLAQPITKTITYPNGDRVEGKITMVRDVTYADGSKGLTTDDKFLNSGTAKYVEHLYYT
WP_003097688.1 -----
WP_033689127.1 -----
WP_020903468.1 -----
WP_000287308.1 -----
WP_033685380.1 -----
WP_033681547.1 -----
WP_023948409.1 -----
WP_021874218.1 -----
KFL97016.1 -----
KFL95230.1 -----

WT-GA
WT-GB1
YP_005861639.1 -----ADKIKANA--
WP_033688740.1 GGAQNDHELYEALQEGMKFNVKTKVEGYALTATVIKLGSKAVDSDPNKTPAGPVNAVRDY
WP_003097688.1 -----DITKIEVTNLAEFGA--FPVGK-----
WP_033689127.1 -----SITKIELTNLAELGAGLAVNT-----
WP_020903468.1 -----DVTKIELTNLAELGSGLAVNS-----
WP_000287308.1 -----SITKIELTNLAELGGGLAVNT-----
WP_033685380.1 -----DTTITIEMKNLTGFGDSFKPGT-----
WP_033681547.1 -----DTTKIELKNLAGFGDAFKAGT-----
WP_023948409.1 -----DTTKIELKNLAGFGDAFKVGT-----
WP_021874218.1 -----NNGQLTLDITVNGSILGNGTVNIKGTVTSNVNSVNSSV---
KFL97016.1 -----NSGNLTLDITINGAVSGNGTVNIKGNVTSNVNENNSLIKGA
KFL95230.1 -----NSGNLTLDITINGAVSGNGTVNIKGNVTSNVNENNSLIKGA

```

Figure A2. Continued.

```

WT-GA
WT-GB1
YP_005861639.1
WP_033688740.1 GDWGSQQLAKAQAED-----KRFKANATKNKTTLANNPTVSVNGVALTLPEYPTETF
WP_003097688.1
WP_033689127.1
WP_020903468.1
WP_000287308.1
WP_033685380.1
WP_033681547.1
WP_023948409.1
WP_021874218.1 ---PTQDQFKAQNYTG-----NRN----NFKNSNIAGNSVNIENGASLTI
KFL97016.1 TADAANAALKDQTSTGTIGTQGTSWASGSSNQNGWTVKGWNYANFSGSKVNVAADANLTI
KFL95230.1 TADAANAALKDQTSTGTIGTQGTSWASGSSNQNGWTVKGWNYANFSGSKVNVAADANLTI

WT-GA
WT-GB1
YP_005861639.1
WP_033688740.1 ANTDAYYSKAAAYNGAVKAYNQIDAQQLSATKLSETGFTYVNVNGKAVPNGSANLYKKI
WP_003097688.1 -----EITAADGT-----VI----
WP_033689127.1 -----EIRATDGT-----VV----
WP_020903468.1 -----EIRATDGT-----VV----
WP_000287308.1 -----EIRETDGT-----VV----
WP_033685380.1 -----KITAADGT-----VI----
WP_033681547.1 -----PIKAADGT-----TI----
WP_023948409.1 -----PIKAADGT-----TI----
WP_021874218.1 NSSEINDGINLTDGGTVRVGDNA--TLNVNLTNASTTATRYHVAGVFAKNGGNFIS----
KFL97016.1 NRSAIGDGIHLANNGTVNVADGGQLTINMNTNNDLNTTARYHNAGIFAVGNGNFTT----
KFL95230.1 NRSAIGDGIHLANNGTVNVADGGQLTINMNTNNDLNTTARYHNAGIFAVGNGNFTT----

WT-GA
WT-GB1
YP_005861639.1
WP_033688740.1 GNERKADVLITAQDTQWSYLRKAGLPTINPDGSEMLTFTSSRDESNVGVTFALSATYNG
WP_003097688.1 -----G
WP_033689127.1 -----G
WP_020903468.1 -----G
WP_000287308.1 -----G
WP_033685380.1 -----G
WP_033681547.1 -----G
WP_023948409.1 -----G
WP_021874218.1 -----G
KFL97016.1 -----G
KFL95230.1 -----G

```

Figure A2. Continued.

WT-GA
WT-GB1
 YP_005861639.1 -----TLYISRNTWGNTTD---GNQPTKVLLSGNVLS-----
 WP_033688740.1 RIVDVNVIAADAEAEAGRTEIVQFETDGTKWEQFMALNLQKDIIDEKTAQGVYPSRQSDVNA
 WP_003097688.1 KVKSIDYKTSTGNNNNKSV--PYWAQRTQRGMTYDERVAEQPAIANETGTYTY-----
 WP_033689127.1 KVKSIDFKTTTGNNNNKSV--PYWAQRTQRGMTYDQRVAEQPAVANETGTYTY-----
 WP_020903468.1 KVKSIDFKTTTGNNNNKSV--PYWAQRTQRGMTYDQRVAEQPAVANETGTYTY-----
 WP_000287308.1 KVKSIDYKTSTGNNNNKST--PYWGQRTQRGMTYDQRVAEQPAVANETGTYTY-----
 WP_033685380.1 EVSSKETTNVSGSRGGES---PFWQSRKKDGKTYEERLAEQPAIPNEVGTTTY-----
 WP_033681547.1 VVKSAAQSNSTGNKGTTT---PYWAQRMKDGMTYEERLAEQPAIPNEVGTTITY-----
 WP_023948409.1 VVKSAAQSNSTGNKGTTT---PYWAQRMKDGMTYEERLAEQPAIPNEVGTTITY-----
 WP_021874218.1 YKSNVNFNTGLGQAIAIGATRPTGTDSDFGGYGARSNDGPTLVQLGDSSTFNFTGRDG
 KFL97016.1 YKSVVTLNTSIGQGIAMTGMRPYVTDTDVFGGYSARDRGDGSQINLGQYSTLNFTGRDG
 KFL95230.1 YKSVVTLNTSIGQGIAMTGMRPYVTDTDVFGGYSARDRGDGSQINLGQYSTLNFTGRDG

WT-GA
WT-GB1
 YP_005861639.1 -----GDTVITISIPSYGIVGVNSPTI-----
 WP_033688740.1 DGRLAEEAGYNPKDWNVDENGNP SAYGTNTFGANYTSMGSKNSLPIALSQNVKTL SMYLN
 WP_003097688.1 -----NIEWNEKADYPNISFGAENL-----
 WP_033689127.1 -----NIEWNDKVKDYPNVSFGASNL-----
 WP_020903468.1 -----NIEWNDKVKDYPNVSFGASNL-----
 WP_000287308.1 -----NIEWNDKVKDYPNVSFGASNL-----
 WP_033685380.1 -----TIRWNDKAKNYAVTTFYAENL-----
 WP_033681547.1 -----TIEWNEKASNYPVTTYSVENL-----
 WP_023948409.1 -----TIEWNEKASNYPVTTYSVENL-----
 WP_021874218.1 IILGNANFISGENSNVHFENKGRGVALDLAANSNI-----
 KFL97016.1 VILGNNSNFNVGDSANVHFENKGRGVALDLAANSNI-----
 KFL95230.1 VILGNNSNFNVGDSANVHFENKGRGVALDLAANSNI-----

WT-GA
WT-GB1
 YP_005861639.1 -----D-----
 WP_033688740.1 SAGAQAAGTIGFMIYDGGDAPQSYGSAQHIIGDFNKEVVKDGVKQTATQPYLGNVKGDP
 WP_003097688.1 -----SG-----
 WP_033689127.1 -----SG-----
 WP_020903468.1 -----SG-----
 WP_000287308.1 -----SG-----
 WP_033685380.1 -----TG-----
 WP_033681547.1 -----TG-----
 WP_023948409.1 -----EISKHSTTYFHSVKGKT-----
 WP_021874218.1 -----NIDDHAVTYFHSVKGKTTNAL
 KFL97016.1 -----NIDDHAVTYFHSVKGKTTNAL
 KFL95230.1 -----NIDDHAVTYFHSVKGKTTNAL

Figure A2. Continued.

```

WT-GA
WT-GB1
YP_005861639.1 -----ANYGSASLKDMGNDKVVIY-----
WP_033688740.1 DFRSTKTDPSGGWVLDDLITSEKYKETPLESGKTVVTTDKGVTGKYLLLPNGNAVIEKSD
WP_003097688.1 -----GGYLAPQISKDTDYKATIKIDGRTVV-----
WP_033689127.1 -----NGYLAPQISKDTPYTATIKIDGRTVL-----
WP_020903468.1 -----DGYLAPEISKDTPYTATIKIDGRTVL-----
WP_000287308.1 -----SGYYAPEISKDTPYTATIKIDGRTVL-----
WP_033685380.1 -----TAIDYYAPNISKDTEYTAAISVNGQPIL-----
WP_033681547.1 -----SGYYAPQISKDTEYTAEIKVDGQKVL-----
WP_023948409.1 -----SGYYAPQISKDTEYTAEIKVDGQKVL-----
WP_021874218.1 ----CTSGSYDGYNYIGVNEGGNITVDEYATFRVIL-----
KFL97016.1 GNTVGASGSFSGYNYIGVNEGGNITVGKFATFRVIL-----
KFL95230.1 GNTVGASGSFSGYNYIGVNEGGNITVGKFATFRVIL-----

```

```

WT-GA
WT-GB1
YP_005861639.1 -----
WP_033688740.1 NTRVLLNQGDVIEMVNPSTKLPIRGIYNHTTGALGEGTLGDEGESQLLDPAVATEYKLRQ
WP_003097688.1 -----EHT-----
WP_033689127.1 -----EHT-----
WP_020903468.1 -----EHT-----
WP_000287308.1 -----EHT-----
WP_033685380.1 -----EHK-----
WP_033681547.1 -----EHT-----
WP_023948409.1 -----EHT-----
WP_021874218.1 -----EGRGDNPWDDVVSLSQNTNTNAAFTSKTGAIVDIRDDNTNFYAEIISFPLGG
KFL97016.1 -----EGRGNNNYDDVVSLSQNTNTNAAFTSKTGAIVDIRDDNTNFYAEIISFPLGA
KFL95230.1 -----EGRGNNNYDDVVSLSQNTNTNAAFTSKTGAIVDIRDDNTNFYAEIISFPLGA

```

```

WT-GA
WT-GB1
YP_005861639.1 -----
WP_033688740.1 AQGNEYVLDGVRANLGVNNDKAYVRGWVDFNNGKFDLNESEIVEVNQNGTYSIKFKNT
WP_003097688.1 -----YTRKGQ-----
WP_033689127.1 -----YTRKGQ-----
WP_020903468.1 -----YTRKGQ-----
WP_000287308.1 -----YTRKGQ-----
WP_033685380.1 -----YTHKAS-----
WP_033681547.1 -----YTHKAT-----
WP_023948409.1 -----YIHKAT-----
WP_021874218.1 SNSRIDIQDPLMLNLQRYSSGGATTGWMPTGGDMINTTSAEYTSNLIYMSGNKGVSFVSG
KFL97016.1 SNTRIDIHDPLMLNLQRYSSGGPTTGWMPIGGDMINTTSNQYTANLIYMSGSKGVFSVDG
KFL95230.1 SNTRIDIHDPLMLNLQRYSSGGPTTGWMPIGGDMINTTSNQYTANLIYMSGSKGVFSVDG

```

Figure A2. Continued.


```

WT-GA
WT-GB1
YP_005861639.1 -----NFTTS
WP_033688740.1 PQLDTSADSLGVRRLRISLTKDEILEPTGVASSGEVEDFETHVIHMPRGTKHETKDFQGR
WP_003097688.1 -----QATFSKQ
WP_033689127.1 -----QPSYQKQ
WP_020903468.1 -----KANYQKQ
WP_000287308.1 -----QPNYQKQ
WP_033685380.1 -----KSAAQKQ
WP_033681547.1 -----KPSVQKQ
WP_023948409.1 -----KPSVQKQ
WP_021874218.1 GDYDPSNPNSSGFVVYQRIKSDGSKQIWLNVNDVNIIPMNGFQTKDIWNQANPDVSITGN
KFL97016.1 TNY-----VVYQKIKSDGSKQIWLNVNGVNIIPMSGFQTKDIWDNQNANPDVSIKGN
KFL95230.1 TNY-----VVYQKIKSDGSKQIWLNVNGVNIIPMSGFQTKDIWDNQNANPDVSIKGN

WT-GA
WT-GB1
YP_005861639.1 GVINPIITIPADNGYGAKPT-----PMQITQPTVKDIT-
WP_033688740.1 EQTVKLPTNAMFTASGKNKDSNYQWAOIENDNLPPKIVLTDKQVASEEAYTPDSELPS
WP_003097688.1 GTSASLRTNNGLGYSNELKSK-----SDTLEINTDSDIR-
WP_033689127.1 GTSASLSENNGLTYLNNEQTGR-----SDSIVLKTDSDIR-
WP_020903468.1 GTSASLSENNGLKYLNEQISR-----SDSIVLKTDSDVR-
WP_000287308.1 GTSASLSENNGLTYLNNEQTGR-----SDSIVLKTDSDVR-
WP_033685380.1 NTSATIMGDNILGYKGSERVRK-----SDAVVINTDSDVR-
WP_033681547.1 STSVSLTLDNGVNYNGSSLVKK-----NDAVVINTDSDVR-
WP_023948409.1 STSVSLTLDNGVNYNGSSLVKK-----NDSVVINTDSDVR-
WP_021874218.1 GLTGGIRANQVHNYNGSPLTGKDAPYYGISTQRASQQIWIPHR--TPLEITGNHTNTIK-
KFL97016.1 DLTSGIRANQVHNYDGTPLTGKDAPYYGISTQRASQQIWFPKH--TQMEVVGSHNTNTIK-
KFL95230.1 DLTSGIRANQVHNYDGTPLTGKDAPYYGISTQRASQQIWFPKH--TQMEVVGSHNTNTIK-

WT-GA
WT-GB1
YP_005861639.1 -----KQY-----
WP_033688740.1 -----MTYKLIL
WP_003097688.1 -----WTINGVE-----QTSAEFHIDVNPVWNPKF--SLSKPNP
WP_033689127.1 NYRKGDDGKVFVERNGATVFTGEYVTVKDKNKVLGKGLKITNPLNNKTEYLLDITYEYD
WP_020903468.1 -----Y-----GVGSKFTIKLP--NSEF--TEFKELE
WP_000287308.1 -----Y-----GVGSKFTIKLP--NADF--TEFKELE
WP_033685380.1 -----Y-----GVGSKFTIKLP--NADF--TEFKELE
WP_033681547.1 -----Y-----GKGSKFTITLP--NDDF--TYFKAID
WP_023948409.1 -----Y-----GKGSKFTIDL--SDEF--TYFRELG
WP_021874218.1 -----YVDEQGNEIFPEN-----TSSLNLKRNIILDITQDQIKKI--QDYALNH
KFL97016.1 -----YVYEDGTPVLDENGNQIVKTQNLNLTRKLTLDITDDKIEEI--QKYALTH
KFL95230.1 -----YVYEDGTPVLDENGNQIVKTQNLNLTRKLTLDITDDKIEEI--QKYALTH

```

Figure A2. Continued.

WT-GA	-----
WT-GB1	NG -----
YP_005861639.1	NSTDQNALKKKMIPNYESI-----YQLAI-----
WP_033688740.1	TAGNEVGTYKVNPAANGKNVSIINGLYETTLTFKPVDAVVGTAAGIAVRAWDDNNSSTGW
WP_003097688.1	GSTNFVNLNTASTLNPN-----KGDSITY-----RLANRWANVKA-----
WP_033689127.1	GSSNFVNLNTASTINPN-----KGDSITY-----RPASRWANVRA-----
WP_020903468.1	GSSNFVNLNTASTINPN-----KGDSITY-----RPASRWANVKA-----
WP_000287308.1	GSSNFVNLNTASTINPN-----KGDSITY-----RPASRWANVKA-----
WP_033685380.1	GSKNTVTNTN-----KADSISY-----RPAGRWNVQA-----
WP_033681547.1	NGKNSI-----KSSEVTY-----RPSNRWSNVQA-----
WP_023948409.1	NGKNSI-----KSSEVTY-----RPSNRWSNVQA-----
WP_021874218.1	TADETLEYIKNSQSVAQDSGWKFTNGSGQTVTDPYATVESPKLDGYTATI-----
KFL97016.1	NADQTLEYIKNAQGVSEDSGWVYTDAQGNTVTDPYATVVSPVEDGYTASI-----
KFL95230.1	NADQTLEYIKNAQGVSEDSGWVYTDAQGNTVTDPYATVVSPVEDGYTASI-----
WT-GA	-----
WT-GB1	-----
YP_005861639.1	-----NETN--GVAPGQDYQNSLPYPSSKIN-----
WP_033688740.1	EATNDTIETSKTSTTLAEKDKVLENTNNGNNGYKSMDTSYIPTVIDVRPVGEDTITEDVQ
WP_003097688.1	-----NENNVWILQDREGGFTLTPLKISPTLELEVTVE--
WP_033689127.1	-----NENNVWILQDGRDSGFTLTPLRISPTLELELTVE--
WP_020903468.1	-----NENNVWILNDGRDSGFTLTPLRISPTLELELTVE--
WP_000287308.1	-----NENNVWILNDGRDSGFTLTPLRISPTLELELTVE--
WP_033685380.1	-----NANNVWILNDGRDNTFTLKATIKSPTELEIEVID--
WP_033681547.1	-----NQNNVWILQDGRDTSFTLKATLKSPTRLELEVID--
WP_023948409.1	-----NQNNVWILQDGRDTSFTLKATLKSPTRLELEVID--
WP_021874218.1	-----QSTNVQGLKVGEDASSVTAKFAVNPSDIVQNGELT
KFL97016.1	-----ESSNVPGITGADGTSVTAKLQYK--EELVQNGELS
KFL95230.1	-----ESSNVPGITGADGTSVTAKLQYK--EELVQNGELS
WT-GA	-----
WT-GB1	-----
YP_005861639.1	-----SAVNYGTVITIPMPKGYV-----LDESATMQLNNFG-----
WP_033688740.1	GKPQSSNPTIPAYATVETVTNDKIEDTKYAANFVILDKTKKPTLATQKENPGKLYTEDTK
WP_003097688.1	-----GYIQEGSNISLSLQSLGVEK-----VIKDKTLTSEFSNVTYNEFGIITG--
WP_033689127.1	-----GAIQEGSTVSMPLQSLGIEK-----VIKDKTLTSEYSNITYENGLIKQG--
WP_020903468.1	-----GAIQEGSVVSMPLQSLGIEK-----VVKDKTLTSEYANITYENGLIKEG--
WP_000287308.1	-----GAIQEGSVVSMPLQSLGIEK-----VIKDKTLTSEYSKITYENGLIKEG--
WP_033685380.1	-----GAIQEDSIVSIGLDKLGVEK-----AITNRTFSDEYSKFIYDEAGRLKD--
WP_033681547.1	-----GVIQEGSTVSMALDKLGIEK-----VLTPTFTSDDFSKIKYDELGRVAK--
WP_023948409.1	-----GVIQEGSTVSMALDKLGIEK-----VLTPTFTSDDFSKIKYDELGRVAK--
WP_021874218.1	DSYKNDGITGI-PDNYVTVVVYKKAKEKGSVKVYHDDTTNTEIPNTEYNTGSVDAGTKV
KFL97016.1	NNYKQNGLSAILPDNYETVVVYKKAKEV-TNTLKFYDDTTKSYISTVADQTATGKENDDV
KFL95230.1	NNYKQNGLSAILPDNYETVVVYKKAKEV-TNTLKFYDDTTKSYISTVADQTATGKENDDV

Figure A2. Continued.

WT-GA
WT-GB1
 YP_005861639.1 -----DKTAI-----
 WP_033688740.1 VDKETTLTLTDGTTATYKPVDKIPANTVIAKDGNEVEVTGTGNVVLNNVRLVTGSQIPAGS
 WP_003097688.1 -----GTAGN-----DKTAATLTVSGGTSI-----
 WP_033689127.1 -----YVGN-----DKTAATLTVSGGESV-----
 WP_020903468.1 -----YVGN-----DKTAATLTVSGGESV-----
 WP_000287308.1 -----YVGN-----DKTAATLTVSGGESV-----
 WP_033685380.1 -----GSVVG-----TDKTAATLTVSGGTPL-----
 WP_033681547.1 -----GSTVGN-----DKTSATLSVTGGTPL-----
 WP_023948409.1 -----GSTVGN-----DKTSATLSVTGGTPL-----
 WP_021874218.1 DYTT-TTTITN-----LENQGYVYVS-TDGTIPSTIE-----
 KFL97016.1 NEKDGASTVKS-----LEDQGYKFINVTDGTPDDTNATVLSGDTFSDV
 KFL95230.1 NEKDGASTVKS-----LEDQGYKFINVTDGTPDDTNATVLSGDTFSDV

WT-GA
WT-GB1
 YP_005861639.1 -----TQDGNIIITVPKSGTQNWNSGG-----P
 WP_033688740.1 KPQSNHPTTVEVQVTLADGTEQTIPAGGTIPGGATIKTPTNTATNTFNNTYNNQGTIP
 WP_003097688.1 -----NGEKEDVTTTVANGWEINASGTPL-----G
 WP_033689127.1 -----NGEKEDVITKVPNGWSIVGDGRVQ-----G
 WP_020903468.1 -----NGEKEDVATTVPNGWSVVGDKVQ-----G
 WP_000287308.1 -----NGEKEDVATTVPNGWSVKGDKVQ-----G
 WP_033685380.1 -----NGGEENVTKVANGWKVEVGAGGN-----S
 WP_033681547.1 -----NGQTEEIATKVNNGWEVSVIPNGS-----T
 WP_023948409.1 -----NGQTEEIATKVNNGWEVSVIPNGS-----T
 WP_021874218.1 -----GNQNVVTVHMKHGVQPVTPDPT-----P
 KFL97016.1 DFGKFGKDGKTFVVHLTHKVPVTPDPT-----P
 KFL95230.1 DFGKFGKDGKTFVVHLTHKVPVTPDPT-----P

WT-GA
WT-GB1
 YP_005861639.1 YQLVGSYNIPMPN-----TATTYAD--AP-----
 WP_033688740.1 AASAGKISSLAKETSADQLVEKGKSITLDGTTYSNNDVIPKGTRTMTTYEDLRNVTLPN
 WP_003097688.1 EPPTGAVVIGLKDLETGKI-----IGHATTKYDGN--RPLNEDGTDNTNVLGKKY---
 WP_033689127.1 EPPTGAVVRTFKDLVTGEV-----IGFEPTRYTGN--IPLSEDGSKDYTNVLGNKY---
 WP_020903468.1 EPPTGAVVRTFKDLVTGEV-----IGFEPTRYTGN--IPLSEDGSKDYTNVLGNKY---
 WP_000287308.1 EPPTGAVVRTFKDLVTGEV-----IGFEPTRYTGN--IPLSEDGSKDYTNVLGNKY---
 WP_033685380.1 QFETGAVTITLVNIETGEE-----FAYEPTSYDGY--AKPNEDGSYETKNILGKKY---
 WP_033681547.1 TVETGAVTTLTKDLETGKI-----IGYIPTEYNGY--APLKEDGTYETKNILGKKY---
 WP_023948409.1 TVETGAVTTLTKDLETGKI-----IGYIPTEYNGY--APLKEDGTYETKNILGKKY---
 WP_021874218.1 DVPKNTPAEAQPDQLTKKVNLTVNYSVSDGSTFTAT--VPANAKQTVTFTGTAYVDKVTG
 KFL97016.1 NVPSNS--KVSDDLTKTATRTIHYVENDQNGAE---LKESTVQTVNYTGTAYVDVVTG
 KFL95230.1 NVPSNS--KVSDDLTKTATRTIHYVENDQNG--AE--LKESTVQTVNYTGTAYVDVVTG

Figure A2. Continued.

WT-GA	-----
WT-GB1	-----
YP_005861639.1	-----
WP_033688740.1	VHIDPQTGEVTSVPRRYTKVTETEIVIENEGTYTLNQDTGEITFIPDKFVGTGTGVTKQ
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----
WP_021874218.1	QLVNA-----TQQNGQWVIDEN-----NTATPQITWTS-----KTSFDKV
KFL97016.1	QMVNAKADGKDAQGNNTYVVDTD-----NKKQPSITWTTDNNGKFAQVTPDAS
KFL95230.1	QMVNAKADGKDAQGNNTYVVDTD-----NKKQPSITWTTDNNGKFAQVTPDAS
 WT-GA	 -----
WT-GB1	----- KTLLKGE -----
YP_005861639.1	-----ITIVQKL-----
WP_033688740.1	QPDVDYNDKVAGDPVTSKYGTDYGGAKYIPIVKPQSKASITRTIHYVYENANDNPTSQDS
WP_003097688.1	-----DVSDYIPPLIKEVDGE-----E
WP_033689127.1	-----DVSNDNVDLVKEVNGE-----E
WP_020903468.1	-----DVSNDPVDLVKEVNGE-----E
WP_000287308.1	-----DVSNDPVDLVKEVNGE-----E
WP_033685380.1	-----DVSNQVPDLTKTIKGV-----E
WP_033681547.1	-----DVTNQIPDLVKPIAGV-----D
WP_023948409.1	-----DVTNQIPDLVKPIAGV-----D
WP_021874218.1	VSPVEQNYHL-ISISDHQDGNDAVATITGL-----TKDSGDITVTVT
KFL97016.1	IKKGDDTWTGTVKSVDEKNAPDVLITGKTTNEDVYVPYTLSQKTYTGTKETKTVTRVIN
KFL95230.1	IKKGDDTWTGTVKSVDEKNAPDVSTITGKTTNEDVYVPYTLSQKTYTGTKETKTVTRVIN
 WT-GA	 -----
WT-GB1	-----
YP_005861639.1	-----
WP_033688740.1	YKDNDPILAI-----
WP_003097688.1	YILAN-----
WP_033689127.1	YILAD-----
WP_020903468.1	YILAD-----
WP_000287308.1	YILAD-----
WP_033685380.1	YIRVD-----
WP_033681547.1	YILAE-----
WP_023948409.1	YILAE-----
WP_021874218.1	YAPNGKIIPV-----DPSGNPI-----
KFL97016.1	YLDNETKQPVSDAVEQTTTLSRTQIKDEKGNVIGYGTVSEDGHSYTLNNDWTIDKNGWVA
KFL95230.1	YLDNETKQPVSDAVEQTTTLSRTQIKDEKGNVIGYGTVSEDGHSYTLNNDWTIDKNGWVA

Figure A2. Continued.

WT-GA	-----
WT-GB1	-----
YP_005861639.1	-----
WP_033688740.1	---DNPVTRTQTIDYTRDYKIFSEAGTTDTAITTTNQVTDASGNIYNVGDITPAGTQF-
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----
WP_021874218.1	---PDAPTPQYPTDP-----TDPSKVTPN-----EPV-PNVPGY-
KFL97016.1	QVSPDETAKGYKETPHFEDGKDASTVAADTPSVTDPQDVTNVVFYDHDTPVTPDKPGHG
KFL95230.1	QVSPDETAKGYKETPHFEDGKDASTVAADTPSVTDPQDVTNVVFYDHDTPVTPDKPGHG
WT-GA	-----
WT-GB1	-----
YP_005861639.1	-----
WP_033688740.1	-----NQGSIIIGKWTASSDQNSKFKEIISPTVKGYTAEVVTADFTPRADG
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----
WP_021874218.1	-----
KFL97016.1	LTHDDLNKDVTRTINYVDTTGAAVNGAPDGKSTYTQTAHFTRTAIVDKVNDKLLGYDING
KFL95230.1	LTHDDLNKDVTRTINYVDTTGAAVNGAPDGKSTYTQTAHFTRTAIVDKVNDKLLGYDING
WT-GA	-----
WT-GB1	-----
YP_005861639.1	-----
WP_033688740.1	KMGHIHNGKQPVGLYTPVADNNKDVGAYEPLVSEVRSDDKDDFDMYVVYKADIQKAKVTY
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----
WP_021874218.1	-----TPSVPTVTPIDPGK-----DTP-----VPYTPE-----
KFL97016.1	DGSVDISPDAGDFAWKSTDANLPAVTSKAPSEVGYSVDTPVVQATTVAYNSEPINVTVT
KFL95230.1	DGSVDISPDAGDFAWKSTDANLPAVTSKAPSEVGYSVDTPVVQATTVAYNSEPINVTVT

Figure A2. Continued.

```

WT-GA
WT-GB1
YP_005861639.1 -----NNDGSQTKTWTGPTVSQDFYGANDQIPLAQVPLY
WP_033688740.1 IDLDATGDARILEVQNANPAPATGADAKTTYGVATLQGKSHTAIPYLTAETIKKYEDRGY
WP_003097688.1 --IPTTGIEGTLSTVNTR-----ARDLYSTEELKENGLENASAYVTPVDYNYVKKTRV
WP_033689127.1 --IPVENARGTLSVTKTR-----ARDLYSEEELKAKGINGSAFVNPVNYDYVKKTKV
WP_020903468.1 --LPAENAKGTLSVTKTR-----ARDLYSEEELKAKGINGSAFVTPAEYDYVKKTKV
WP_000287308.1 --IPAENTKGTLSVTKTR-----ARDLYSEEELKAKGINGSAFVTPAEYDYVKKTKV
WP_033685380.1 --VPSKGTGTGNIGPKR-----ASAIYSSEELQANGVNPNAFVNNVVYYYVKKTKV
WP_033681547.1 --IPPKGPAGTINVSDKR-----IRDIYTADEITAAGLNPNAVNNVIYSYVKKTKV
WP_023948409.1 --IPPKGPAGTINVSDKR-----IRDIYTADEITAAGLNPNAVNNVIYSYVKKTKV
WP_021874218.1 --TPAKDQKAVVNYVDAD-----EDNKLITSSGDLTGKAGTKIDYSTNSTIEDLTNKG
KFL97016.1 YSKNAQQGSFQIHYIDED-----NNAILHQDTVSDKIGDSVTYSTADQIQLWESKGY
KFL95230.1 YSKNAQQGSFQIHYIDED-----N-NNAILHQDTVSDKIGDSVTYSTADQIQLWESKGY

```

```

WT-GA
WT-GB1
YP_005861639.1 AKA-----
WP_033688740.1 ELVTDDYTNNQTGTAIEGGRKFDDDKQAFNVYLRH-----
WP_003097688.1 EEV-----NRTIKYVYA-----
WP_033689127.1 EEV-----NRTIKFVYA-----
WP_020903468.1 EEV-----NRTIKFVYA-----
WP_000287308.1 EEV-----NRTIKFVYA-----
WP_033685380.1 EEV-----NRTIKYVYA-----
WP_033681547.1 EEV-----NRTIKYVYA-----
WP_023948409.1 EEV-----NRTIKYVYA-----
WP_021874218.1 VLVNDGFPKDAT-----YDNDNTTQTYTVVLRH-----
KFL97016.1 VLDQDGYTTQTT-----VNEDNNGKTYIVSFKHGRKNGTTETLVPTETIHFQYAD
KFL95230.1 VLDQDGYTTQTT-----VNED-NNGKTYIVSFKHGRKNGTTETLVPTETIHFQYAD

```

```

WT-GA
WT-GB1
YP_005861639.1 -----
WP_033688740.1 -----
WP_003097688.1 -----
WP_033689127.1 -----
WP_020903468.1 -----
WP_000287308.1 -----
WP_033685380.1 -----
WP_033681547.1 -----
WP_023948409.1 -----
WP_021874218.1 -----
KFL97016.1 GTKAADDVHGNAGDFKFTRTPIIDTVTGQIVDPGTWNKESYTFDDGQKNVKVINGYVADK
KFL95230.1 GTKAADDVHGNAGDFKFTRTPIIDTVTGQIVDPGTWNKESYTFDDGQKNVKVINGYVADK

```

Figure A2. Continued.

WT-GA	-----
WT-GB1	-----
YP_005861639.1	-----
WP_033688740.1	-----KKVTRKIKDTQE
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----
WP_021874218.1	-----GTQPV-----NPTNPGK-----
KFL97016.1	ATYGNKTATPTDLNVEDTVTYRKISNIIIPVDENGNQIPGTTTPVDYKNDPSDPTKVTPDEE
KFL95230.1	ATYGNKTATPTDLNVEDTVTYRKISNIIIPVDENGNQIPGTTTPVDYKNDPSDPTKVTPDEE
WT-GA	-----
WT-GB1	-----
YP_005861639.1	-----AY-----
WP_033688740.1	VRTTIEYKYASTDDVPA-----
WP_003097688.1	-----DDAQG-----
WP_033689127.1	-----DNVEG-----
WP_020903468.1	-----DNVAG-----
WP_000287308.1	-----DNVAG-----
WP_033685380.1	-----DDVKD-----
WP_033681547.1	-----DDVKN-----
WP_023948409.1	-----DDVKN-----
WP_021874218.1	-----PGEPINPNPDG-----PKYPTG-----
KFL97016.1	SPKVP SGWTISP NQPEGVTPNTTTNTAKVTPVDPTKPTNVVYTKDNAPVDKATVIVRYHD
KFL95230.1	SPKVP SGWTISP NQPEGVTPNTTTNTAKVTPVDPTKPTNVVYTKDNAPVDKATVIVRYHD
WT-GA	-----
WT-GB1	-----
YP_005861639.1	-----
WP_033688740.1	-----
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----
WP_021874218.1	-----
KFL97016.1	DTTNLDLPESFDSGNKEVGTDGTGYTQADINKVVQEYEAKGYYYVTTDGTLPPTTIPAGGAT
KFL95230.1	DTTNLDLPESFDSGNKEVGTDGTGYTQADINKVVQEYEAKGYYYVTTDGTLPPTTIPAGGAT

Figure A2. Continued.

```

WT-GA
WT-GB1
YP_005861639.1 -----GGNQLLNNGHKQIVAYFGA--
WP_033688740.1 -----DKRGTTAAPTVTETLHFERDRT
WP_003097688.1 -----LAGQQVFDPTTQTVSYKGT--
WP_033689127.1 -----LAGTEVFPSQKQTVSYTGS--
WP_020903468.1 -----LAGTEVFPSQKQTVSYTGS--
WP_000287308.1 -----LAGTEVFPSQKQTVSYTGS--
WP_033685380.1 -----LAGQEVFEPTKQTVSYTGT--
WP_033681547.1 -----LAGQQVFEPTKQTVSYTGT--
WP_023948409.1 -----LAGQQVFEPTKQTVSYTGT--
WP_021874218.1 -----SNEVTKTVTRTIQYLDEDEGNKVSDSVEQPVNFTASGV
KFL97016.1 IVVHLAHNQIPVGPDPDPKHGVDPDQVKKAYTSTLHYQDSEGKTLSPDQQQTSTWTRTVT
KFL95230.1 IVVHLAHNQIPVGPDPDPKHGVDPDQVKKAYTSTLHYQDSEGKTLSPDQQQTSTWTRTVT

WT-GA
WT-GB1
YP_005861639.1 -----
WP_033688740.1 IDYTLAAKEYPTEYAAYKAVLDASGYDSPEEYKARVVYYDHIITNKAIAADATDAQKAIV-
WP_003097688.1 -----VKT-
WP_033689127.1 -----IKL-
WP_020903468.1 -----IKL-
WP_000287308.1 -----IKL-
WP_033685380.1 -----IQV-
WP_033681547.1 -----VKL-
WP_023948409.1 -----VKL-
WP_021874218.1 LDKVT-----GEWTTPLTWSV---DQTVSAVKSPVVSGYHLVSVDRDQDGNVVKDVTL-
KFL97016.1 VDTVNTQIVNGGKYDT--NWTLQDANDKYSNFTVPVVEGY--VARKTTNNGATVTTTVAG
KFL95230.1 VDTVNTQIVNGGKYDT--NWTLQDANDKYSNFTVPVVEGY--VARKTTNNGATVTTTVAG

WT-GA
WT-GB1
YP_005861639.1 -----TNESIASYSNYSSNFTFNFDESLG---VTELKTPTI
WP_033688740.1 -----TFGPWTPVGGTSNDAITLSDAEKAKDDKFNLVNSPEV
WP_003097688.1 -----NAEGKAEIGSNDKPIYINWKGIGDTN--LPEVTVPOK
WP_033689127.1 -----TAEGKAVINSNDRPVYINWKGTDGQSTDLPPELAVPOK
WP_020903468.1 -----TAEGKAVINSNDRPVYINWKGTDGQSTDLPPELAVPOK
WP_000287308.1 -----TAEGKAVINSNDRPVYINWKGTDGQSTDLPPELAVPOK
WP_033685380.1 -----NDKNEAQVDANRKPIYINWVGNGDTN--LPEVTVPOK
WP_033681547.1 -----NSDGKAAVDSNDKPIYVNWVGNGDTN--LPEVTVPOK
WP_023948409.1 -----NSDGKAAVDSNDKPIYVNWVGNGDTN--LPEVTVPOK
WP_021874218.1 -THDDNSYIVTVRYAKNGKIIPVDPNGH-PIPNVPQPYPTDPNNPAKV--TPDEPVPI
KFL97016.1 QTKVQQNLEDTVVYDKVGKLVPGPDGKTPIPDAPTPSYPNPDPTDPTKV--IPNEPVPDV
KFL95230.1 QTKVQQNLEDTVVYDKVGKLVPGPDGKTPIPDAPTPSYPNPDPTDPTKV--IPNEPVPDV

```

Figure A2. Continued.


```

WT-GA
WT-GB1
YP_005861639.1 PGT-----SNKYTITYA-----DGTTSSEQV---NAG
WP_033688740.1 TGYVPDNAT-VEATAAIDAEADDYKITVLYTPVAQKAVVKFVEVDPTNTDKVITPGLADP
WP_003097688.1 EGYIASVEK-V-PVQPTTATDEDEYEVVVTYSPI-QKAKTTFVYQDKDGNVKQV----EGN
WP_033689127.1 EGYIASVEK-V-PVQATTATDEDEYEVVVKYTAI-QKAKTTFV--DEKGNA--I----PGV
WP_020903468.1 EGYIASVEK-V-PVQATTATDEDEYEVVVKYTAI-QKAKTTFV--DEKGNA--I----PGV
WP_000287308.1 EGYIASVEK--VPVQATTATDEDEYEVVVKYTAI-QKAKTIFV--DEKGNAI-----PGV
WP_033685380.1 EGYIASVEK--VPVQPTTATDEDEYEVVVTYTAI-QKAKTTFV--DEKGNPI-----PGV
WP_033681547.1 EGYIASVEK--VPVQPTTATDEDEYEVVVTYSPI-QKAKTTFVYQDKDGNVKQV----EGN
WP_023948409.1 EGYIASVEK--VPVQPTTATDEDEYEVVVTYSPI-QKAKTTFVYQDKDGNVKQV----EGN
WP_021874218.1 PGMTPSVPT-VTPDPGKDTFVPYTPVAPAKD---QVAQVIYRDVNDPNKVTQL---ATS
KFL97016.1 PGYTPVDPTPITPEDPTKDTFVPYTK----DPVKAGLTVQYIDQDNNSV-----IKS
KFL95230.1 PGYTPVDPTPITPEDPTKDTFVPYTK----DPVKAGLTVQYIDQDNNSV-----IKS

WT-GA
WT-GB1
YP_005861639.1 NTITGTGVITNIV----VSPDNFERDQSTAVDLPTNKF-----ANQTTQSVNAFE
WP_033688740.1 IAVTGKSEAAYPATTATSVTDKIAELVKKGyelVDNGF--VSADKFDKDAVDQEVVVKF
WP_003097688.1 TPisetGKGgDKLTkadeIAAKIkeAQNKgyELVSNTY--PTDGAfDKDVNTDQEFVTl
WP_033689127.1 AEITEQGGSETPLTKEADVKAkIAELenKgyELVSNTY--PEGgKFDTDKDtdQEFKvIl
WP_020903468.1 AEITEQGGSETPLTKEADVKAkIAELenKgyELVSNTY--PEGgKFDTDKDtdQEFKvIl
WP_000287308.1 AEITEQGGSETPLTKEADVKAkIAELenKgyELVSNTY--PEGgKFDTDKDtdQEFKvIl
WP_033685380.1 DEITEQGGSEAPLTKEADVKAkIAELenKgyELVSNTY--PEGgKFdKEAGVDQEFKvTL
WP_033681547.1 TPisetGKGgDKLTkadeIAARIkeAQNKgyEVVSNTY--PTDGVfDKDvDtdQEFVTl
WP_023948409.1 TPisetGKGgDKLTkadeIAARIkeAQNKgyEVVSNTY--PTDGVfDKDvDtdQEFVTl
WP_021874218.1 GDLTGKAGSEIDY----NAQSEIDNlNKgyVLKNNGF--PAGAVfDNDdNKtQTfYIDf
KFL97016.1 DAVDGNIGDKIDY----STASSITDFENKgyILVTDGfTGAGDEfTTENN-GQVYKVVF
KFL95230.1 DAVDGNIGDKIDY----STASSITDFENKgyILVTDGfTGAGDEfTTENN-GQVYKVVF

WT-GA
WT-GB1
YP_005861639.1 AYGAVPETV-----KSGTQLVANLTFTG-----TIQQGNITKYLTSKI--
WP_033688740.1 KAKVVDVPSFDPTKpASNDNPKPTPGVTPIDPNNPDGPKWTEALINAVKVQEEVTRTIKY
WP_003097688.1 KERVVPVTPDQPK-----TSG-TPVDPNNPDGPKYPA-GLEEKDLNKIVTRTITY
WP_033689127.1 KQKEVTVTPDQPK-----TPG-TPVDPNNPDGPKYPA-GLEEKDLNKIVTRTITY
WP_020903468.1 KQKEVTVTPDQPK-----TPG-TPVDPNNPDGPKYPA-GLEEKDLNKIVTRTITY
WP_000287308.1 KQKEVTVTPDQPK-----TPG-TPVDLNNDGPKYPA-GLEEKDLNKIVTRTITY
WP_033685380.1 KERVVPVTPDQPK-----TPG-TPVDPNNPDGPKYPA-GLEEKDLNKIVTRTITY
WP_033681547.1 KERVVPVTPDQPK-----TPG-TPVDPNNPDGPKYPA-GLEEKDLNKIVTRTITY
WP_023948409.1 KERVVPVTPDQPK-----TPG-TPVDPNNPDGPKYPA-GLEEKDLNKIVTRTITY
WP_021874218.1 VHGTVPVTPDTPG-----KPG-EPINPNDPDGPKWPD-GTSEdSLKSGTQTIHY
KFL97016.1 KHGTRPVTPENPA-----DPN-EPVDPDHPDTPTPSNPNLSKEDLQKTITRTIEY
KFL95230.1 KHGTRPVTPENPA-----DPN-EPVDPDHPDTPTPSNPNLSKEDLQKTITRTIEY

```

Figure A2. Continued.

```

WT-GA
WT-GB1
YP_005861639.1 -----QVDQTVVSSADLTSSSGI-----FGYQ
WP_033688740.1 VYEDGTPVAESDLTSVADKKVKTLKFTRSGKINVATGEITYGD-----W-SA
WP_003097688.1 VYEDGTPVLNED--GTPKTVTQEAKFTREAKVNLVTGEVITYGD-----W-TP
WP_033689127.1 VYADGTPVLNED--GTPKTVTQEAKFTREAKVNLVTGDTVITYGD-----W-SE
WP_020903468.1 VYEDGTPVLNED--GTPKTVTQEAKFTREAKVNLVTGEVITYGD-----W-SE
WP_000287308.1 VYEDGTPVLNED--GTPKTVTQEAKFTREAKVNLVTGEVITYGD-----W-SE
WP_033685380.1 VYEDGTPVLSED--GTPKTVTQEAKFTREAKVNLVSGEVITYGD-----W-SE
WP_033681547.1 VYEDGTPVLNED--GTPKTVTQEAKFTREAKVNLVTGEVITYGD-----W-TP
WP_023948409.1 VYEDGTPVLNED--GTPKTVTQEAKFTREAKVNLVTGEVITYGD-----W-TP
WP_021874218.1 VYSDGSKAKDDN-----VQSFDFTKSAVVVKVTGEII-----SQTG--WNVD
KFL97016.1 KYADGTQAHELV-----KQELTFTGKGTIDLVTGNLVTVDEDGNITSQNGKITWNHE
KFL95230.1 KYADGTQAHEPV-----KQELTFTGKGTIDLVTGNLVTVDEDGNITSQNGKITWNHE

```

```

WT-GA
WT-GB1
YP_005861639.1 SNTATGQSN---VGYL-----
WP_033688740.1 DQTFEAVTSPTLEKYTAAVAGITPAVADVPAKTVAATDKDFEETVIYSTKPTTVDPNKPT
WP_003097688.1 AQDLSEVTSPVVKGYL-----AD-----
WP_033689127.1 AKDLAEVKSPVVTGFL-----AD-----
WP_020903468.1 AKDLAEVKSPVVTGFL-----AD-----
WP_000287308.1 AKDLAEVKSPVVTGFL-----AD-----
WP_033685380.1 AKDLPEVKSPKVDGYL-----AD-----
WP_033681547.1 AQDLAEVKSPVVKGYL-----AD-----
WP_023948409.1 AQDLAEVKSPVVKGFL-----AD-----
WP_021874218.1 SHTFGNVDTFVIDGYH-----AD-----
KFL97016.1 SQEFEAVPAIDHDGYYISSINQSNSTASVDGQ-----
KFL95230.1 SQEFEAVPAIDHDGYYISSINQSNSTASVDGQ-----

```

```

WT-GA
WT-GB1
YP_005861639.1 -----
WP_033688740.1 DPTNPNVTPQPDDVVPNDPKGRITYRELGLIEEVTHTVHYKLEDGSDAGIADNVQTLTFTR
WP_003097688.1 -----
WP_033689127.1 -----
WP_020903468.1 -----
WP_000287308.1 -----
WP_033685380.1 -----
WP_033681547.1 -----
WP_023948409.1 -----
WP_021874218.1 -----
KFL97016.1 -----
KFL95230.1 -----

```

Figure A2. Continued.

```

WT-GA
WT-GB1
YP_005861639.1 -----SVYAGG----GQTNNIYEP
WP_033688740.1 TAEVDPVTGAISNFGTWKAKGGDTTIDAVTTPNKDGYVASAKTSTERTNVAATDKDSEET
WP_003097688.1 -----KATVPTVNVVTADSKDTTEV
WP_033689127.1 -----KASVPVVNVVTGDSKDITEV
WP_020903468.1 -----KASVPVVNVVTGDSKDITEV
WP_000287308.1 -----KASVPVVNVVTGDSKDITEV
WP_033685380.1 -----KASVAVVNVVTGDSEDIKEV
WP_033681547.1 -----KATVPTTKVTADSENTTEV
WP_023948409.1 -----KASVPVVNVVTGDSEDIKEV
WP_021874218.1 -----KRTAGGTTITPDDLNKEVT
KFL97016.1 -----TGAVGTETVTTPNSQNGNIV
KFL95230.1 -----TGAVGTETVTTPNSQNGNIV

```

```

WT-GA
WT-GB1
YP_005861639.1 IFYYVLPEWFSVY-----
WP_033688740.1 IIYRKLGSYVPVPEGITP-----
WP_003097688.1 VTYKPLGSWVPNIPGQPTN-----
WP_033689127.1 VTYKPLGSWIPNIPGKTPT-----
WP_020903468.1 VTYKPIGSWIPNIPGQPTN-----
WP_000287308.1 VTYKPIGSWIPNIPGQPTN-----
WP_033685380.1 VTYKPLGSWVPNIPGQPTD-----
WP_033681547.1 VTYKPLGSWIPNIPGQPTD-----
WP_023948409.1 VTYKPLGSWVPNIPGQPTN-----
WP_021874218.1 VTYTPNGKIIPVDPNGN-----
KFL97016.1 ITLTRNPD-VPVAAQGSINYIDDTTGQTIESANFSGNVGQKINYTTAGSIKNWEAKGYNL
KFL95230.1 ITLTRNPD-VPVAAQGSINYIDDTTGQTIESANFSGNVGQKINYTTAGSIKNWEAKGYNL

```

```

WT-GA
WT-GB1
YP_005861639.1 -----DFSTDYTK-----
WP_033688740.1 -----PADADLNPKYPNATPADPTKPGTP-TETPVV
WP_003097688.1 -----PI-----KYPNDPQDPTKPGQPTETL
WP_033689127.1 -----PI-----KYPNNPDPTKPG-DKPIL
WP_020903468.1 -----PI-----KYPNNPDPTQPGKPTEVL
WP_000287308.1 -----PI-----KYPNNPDPTQPGKPTEVL
WP_033685380.1 -----PI-----KYPNDPTDPTTPGTDKPKV
WP_033681547.1 -----PI-----KYPNDPTDPTKPGDKPKVL
WP_023948409.1 -----PI-----KYPNDPTDPTKPGQPTTEVV
WP_021874218.1 -----PI-PNV-----TPQYPTDPTDPTKV-TPDEPV
KFL97016.1 VSNNFKDGEEVFDTGKNAFEVHLVHATTPVTPENPGKPGEPVNPTNPDDPHKY--PDNYV
KFL95230.1 VSNNFKDGEEVFDTGKNAFEVHLVHATTPVTPENPGKPGEPVNPTNPDDPHKY--PDNYV

```

Figure A2. Continued.

WT-GA	----- YYIKLINNAKT
WT-GB1	----- QYAND -----
YP_005861639.1	----LPDFVP-----NTNNGVIGTPKLSVFTVPTE-----
WP_033688740.1	PY--IPGTTTPVGPNKGKPLTPKDPNDPTKGYEVPDLPTDPTENTTI----TYVKDGSQVAV
WP_003097688.1	PY--VPGFTPKDKDGNPLKPVNPNNPEEGYIVPDLPTDPSQDTPI----NYVKDTQKAKT
WP_033689127.1	PY--EPGMTPKDGNQPLKPVDPSPDPTKGYIVPDLPTDPSQDTPI----NYVKDTQKAKT
WP_020903468.1	PY--VPGFTPKDKDGNPLKPVDPDPTKGYEVPNLPTDPSQDTPI----NYVKDMQKAKT
WP_000287308.1	PY--VPGFTPKDKDGNPLKPVDPADPTKGYIVPDLPTDPSQDTPI----NYVKDTQKAKT
WP_033685380.1	PY--VEGFTPKDKDGNPLKPVDPNDPKEGYEVPNLPTDPSQDTPI----NYVKDTQKAKT
WP_033681547.1	PY--VPGMTPKDKDGNPLKPVDPNDPTKGYEVPNVPTNPGEDTPI----NYVKDTQKAKT
WP_023948409.1	PY--VPGYTPKDGNGQPLKPVDPNNPTKGYEVPSVPTNPGEDTPI----NYVKDTQKAKT
WP_021874218.1	PN--IPGLTP-----SVPTVTPDTPGKDTVPVYNPVVPAKDQAAVV--NYVDAEDNKL
KFL97016.1	PQELAKTVTR-----DVTYVYADGSQAEAPVHQEVKFTGNGYLDLVTGEYVTVDNNGKI
KFL95230.1	PQELAKTVTR-----DVTYVYADGSQAEAPVHQEVKFTGNGYLDLVTGEYVTVDNNGKI
 WT-GA	 V-----EGVE-----
WT-GB1	-----NGVD-----
YP_005861639.1	-----ISGLS-----
WP_033688740.1	T-----HFIEVNSETDKTEKGAV-----
WP_003097688.1	T-----FVDEKGNPIPGVD-----
WP_033689127.1	T-----FVDEKGNPIPGVA-----
WP_020903468.1	T-----FVDEKGNPIPGVD-----
WP_000287308.1	T-----FVDEKGNPIPGVD-----
WP_033685380.1	T-----FVDEKGNPIPGVD-----
WP_033681547.1	T-----FVDEKGNPIPGVD-----
WP_023948409.1	T-----FVDEKGNPIPGVD-----
WP_021874218.1	I-----TSSGDLTGKA-----
KFL97016.1	TGKGQINWTPESANFDTAKSIDTSKYQIVGIKENNTTANVDQTTGVVAGETVTQNSNNSS
KFL95230.1	TGKGQINWTPESANFDTAKSIDTSKYQIVGIKENNTTANVDQTTGVVAGETVTQNSNNSS
 WT-GA	 -----
WT-GB1	-----
YP_005861639.1	-----RQVVKIDYSG
WP_033688740.1	-----AESVVDTGDTGKAFTKAADVTATIEALKA
WP_003097688.1	-----AITEEGSDTPLTKEAEVKAKIKELEN
WP_033689127.1	-----EITEQGSDTPLTKEAEVKAKIKELEN
WP_020903468.1	-----AITEQGSDTPLTKEADVKAKELEN
WP_000287308.1	-----AITEEGSDTPLTKEADVKAKELEN
WP_033685380.1	-----AITEEGSDTPLTKEAEVKAKIKELEN
WP_033681547.1	-----AITEEGSDTPLTKEAEVKAKIKELEN
WP_023948409.1	-----AITEEGSDTPLTKEADVKAKELEN
WP_021874218.1	-----GETINYSTADTIKDLEN
KFL97016.1	VVITLANKPAPVVEKGSITVKVHDLTDNVDLPQYGKESGEQEVGTSFTYDKNNAVITELIN
KFL95230.1	VVITLANKPAPVVEKGSITVKVHDLTDNVDLPQYGKESGEQEVGTSFTYDKNNAVITELIN

Figure A2. Continued.

```

WT-GA
WT-GB1
YP_005861639.1  TGYNFLAG---QGANNQIHLNLTLPDGTNGTYQGIIYIVSPTTK-LTNTAYNPNNTSNFA-
WP_033688740.1  KGYTVVENNYPTDGTFDADSKTNQVYKVLVTAKPITVNPNDPTPTKGQPIDPNNPTGPKW
WP_003097688.1  KGYELVSNTYPEGGKFDKDKGTDQEFKVTLKAKEVTVTPDQPK-TPGTFPVDPNNPDGPKY
WP_033689127.1  KGYELVSNTYPEGGKFDKDKDTDQEFKVTLKAKEVTVTPDQPK-TPGTFPVDPNNPDGPKY
WP_020903468.1  KGYELVSNTYPEGGKFDKDKDTDQEFKVTLKAKEVTVTPDQPK-TPGTFPVDPNNPDGPKY
WP_000287308.1  KGYELVSNTYPEGGKFDKDKDTDQEFKVTLKAKEVTVTPDQPK-TPGTFPVDPNNPDGPKY
WP_033685380.1  KGYELVSNTYPEGGKFDKDKDTDQEFKVTLKERVVPVTPDQPK-TPGTFPVDPNNPDGPKY
WP_033681547.1  KGYELVSNTYPEGGKFDKDKDTDQEFKVTLKERVVPVTPDQPK-TPGTFPVDPNNPDGPKY
WP_023948409.1  KGYELVSNTYPEGGKFDKDKDTDQEFKVTLKERVVPVTPDQPK-TPGAPVDPNNPDGPKY
WP_021874218.1  KGYVLVNDGFPAGAKYDSDNTTQIYTVVLKHGTTTITPDKPG-KPGEFIPNPNPDGPKW
KFL97016.1      KGYKLVDDGENVPSEVAKGAKT---ITILVEHDTVPTPENPG-KPGEFIPNPNPDGPKW
KFL95230.1      KGYKLVDDGENVPSEVAKGAKT---ITILVEHDTVPTPENPG-KPGEFIPNPNPDGPKW

WT-GA
WT-GB1
YP_005861639.1  PSGITFNPDWVQG--NTSNLYYVGANGFTINQ-----
WP_033688740.1  TPELIKELEDGRTEEVKRTIKYVYADGSKAADSVQETKEFKRSATINPVTGKVTFGDWSP
WP_003097688.1  PAGL-EEKDLNKT--VTRTITYYADGTPVLN-----
WP_033689127.1  PAGL-EEKDLNKT--VTRTITYYADGTPVM-----
WP_020903468.1  PAGL-EEKDLNKT--VTRTITYYVEDGTPVLN-----
WP_000287308.1  PAGL-EEKDLNKT--VTRTITYYVADGTPVLN-----
WP_033685380.1  PAGL-EEKDLNKT--VTRTITYLYEDGTPVLN-----
WP_033681547.1  PAGL-EEKDLNKT--VTRTITYYVEDGTPVLN-----
WP_023948409.1  PAGL-EEKDLNKT--VTRTITYYVEDGTPVLN-----
WP_021874218.1  PDNS-GENNLSKT--GTQTIHYTGA-----
KFL97016.1      PEGT-DENSVKRT--GTQTIHYEGA-----
KFL95230.1      PEGT-DENSVKRT--GTQTIHYEGA-----

WT-GA
WT-GB1
YP_005861639.1  AQTFEAVTSPKVTNFTPDKESVPAAEVTATAEDINETVIYTTKPANIDPSKPTDPNTPNV
WP_033688740.1  -----
WP_003097688.1  -----
WP_033689127.1  -----
WP_020903468.1  -----
WP_000287308.1  -----
WP_033685380.1  -----
WP_033681547.1  -----
WP_023948409.1  -----
WP_021874218.1  -----
KFL97016.1      -----
KFL95230.1      -----

```

Figure A2. Continued.

```

WT-GA
WT-GB1
YP_005861639.1 -----VGGANT-----ASVAQGNQNDILV-----
WP_033688740.1 TPRPDDRVPNDPKGRITYKELGLIEEVTHTVHYKLADGSDAGIPDNVQTLTFTRTADLDPV
WP_003097688.1 ----EDGTPK-----TVTQEAKFTREAKVNLV
WP_033689127.1 ----ENGAPK-----VVTQEAKFTREAKVNLV
WP_020903468.1 ----EDGTPK-----TVTQEAKFTREAKVNLV
WP_000287308.1 ----EDGTPK-----TVTQEAKFTREAKVNLV
WP_033685380.1 ----EDGTPK-----VVTQEAKFTREAKVNLV
WP_033681547.1 ----EDGTPK-----VVTQEAKFTREAKINLV
WP_023948409.1 ----EDGTPK-----TVTQEAKFTREAKVNLV
WP_021874218.1 ----GDKTPE-----DNKQEFTFTKTMVVDNV
KFL97016.1 ----GDKTPS-----DDVQTFDFTKKMLVDKV
KFL95230.1 ----GDKTPS-----DDVQAFDFTKKMLVDKV

```

```

WT-GA
WT-GB1
YP_005861639.1 -----DSGVSDLYGSN-----
WP_033688740.1 TGAISNFGTWTAKDNDTTIDAITTPNKPGYVASAAKSTERTNVQATDKDSEETIIYRKLK
WP_003097688.1 TGDV-TYGDWT-PAQD-----
WP_033689127.1 TGEV-TYGDWS-EAKD-----
WP_020903468.1 TGEV-TYGDWS-EAKD-----
WP_000287308.1 TGEV-TYGDWS-EAKD-----
WP_033685380.1 TGEV-TYGDWT-PAQD-----
WP_033681547.1 TGEV-TYGDWT-PAQD-----
WP_023948409.1 TGEV-TYGDWT-PAQD-----
WP_021874218.1 TGKVIITDGAWNVTSH-----
KFL97016.1 TGKIIDSGEWNVTSH-----
KFL95230.1 TGKIIDSGEWNVTSH-----

```

```

WT-GA
WT-GB1
YP_005861639.1 -----
WP_033688740.1 SYVPVIEGVTPPTGTDLTPKPYENPINEDPTRPGTPTETPVVPYIPGTTTPVGPNKPLT
WP_003097688.1 -----
WP_033689127.1 -----
WP_020903468.1 -----
WP_000287308.1 -----
WP_033685380.1 -----
WP_033681547.1 -----
WP_023948409.1 -----
WP_021874218.1 -----
KFL97016.1 -----
KFL95230.1 -----

```

Figure A2. Continued.

WT-GA	-----
WT-GB1	-----
YP_005861639.1	-----
WP_033688740.1	PKDPNDPTKGYEVKVPEDPTQNTTITYVKDGSQVALVHFIAKADGTAVHVSVAEAGDTGK
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----
WP_021874218.1	-----
KFL97016.1	-----
KFL95230.1	-----

WT-GA	-----
WT-GB1	-----
YP_005861639.1	-----
WP_033688740.1	AIKTTNIDNVKAELEAKGYEVVAPTDAAYTAERVAFYAEANRTFDDKDDKGNDGISQVYY
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----
WP_021874218.1	-----
KFL97016.1	-----
KFL95230.1	-----

WT-GA	-----
WT-GB1	-----
YP_005861639.1	-----
WP_033688740.1	VIVKEGITPIDPKPLDPNTPDVTPKPGDKVPGDPKQRTFEQLGLLDEVNRTINYRYANT
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----
WP_021874218.1	-----
KFL97016.1	-----
KFL95230.1	-----

Figure A2. Continued.

WT-GA	----- SLKNEILKALPT
WT-GB1	-----
YP_005861639.1	-----EMEYAVRLINGSGT
WP_033688740.1	DKVDADKRGQEARPTVEQKLYSRKGNLNKVTGEITYTSDWTKPQILAEVTSPVIEGYVA
WP_003097688.1	-----LAEVKSPVVKGFLLA
WP_033689127.1	-----LAEVKSPVVTGFLA
WP_020903468.1	-----LPEVKSPVVKGYLA
WP_000287308.1	-----LPEVKSPVVKGYLA
WP_033685380.1	-----LAEVKSPVVPGYLA
WP_033681547.1	-----LAEVKSPVVKGYLA
WP_023948409.1	-----LAEVKSPVVKGYLA
WP_021874218.1	-----FGNVDTFVIDGYHA
KFL97016.1	-----FGYKDTFVIDGYHA
KFL95230.1	-----FGYKDTFVIDGYHA
WT-GA	E -----
WT-GB1	----- YDDATKTFVTE -----
YP_005861639.1	K-LTNVAMVNLPQASDTSF--AFQLNGRPVY----NGDKTGYTFLYSTELGNLKSNN
WP_033688740.1	DIKAAEKVENVAHDAADSVVNVVYTPLGKYVPKVPEGFEVPKVEKPQYPNDPTDPTKPGT
WP_003097688.1	D-KASVAVVNVTDGSEDIKEVVYTKPLGSWVP---NIPGQPTSPIKYPNDPTDPTKPGQ
WP_033689127.1	D-KASIPVVNVTDGSKDITEVVYTKPIGSWIP---NIPGQPTTPIKYPNDPQDPTKPGQ
WP_020903468.1	D-KATVPATKVTADSENTKEVVYTKPIGSWIP---NIPGQPTNPIKYPNDPTDPTKPGQ
WP_000287308.1	D-KATVPATKVTADSENTKEVVYTKPIGSWIP---NIPGQPTNPIKYPNDPTDPTKPGQ
WP_033685380.1	D-KASVPVVNVTDGSKDTEVVYTKPLGSWVP---NIPGQPTNPIKYPNDPTDPTKPGS
WP_033681547.1	D-KATVPPTKVTADSENTTEVVYTKPLGSWVP---NIPGQPTDPIKYPNDPTDPTKPGQ
WP_023948409.1	D-KVTVPPTKVTADSENTTEVVYTKPLGSWIP---NIPGQPTTPIKYPNDPTDPTKPGK
WP_021874218.1	D-KRTAGGTTITPDDLNKTVTVNYTPNGKIIIPVDPNGNPIPNVPTPQYPTDPTDPTKV-T
KFL97016.1	D-KRNAGGSVVTPNDLNKKVVVYTKPNGKIIPTDPSGNPIPNVPTPTYPTDPTDPTKV-V
KFL95230.1	D-KRNAGGSVVTPNDLNKKVVVYTKPNGKIIPTDPSGNPIPNVPTPTYPTDPTDPTKV-V
WT-GA	-----
WT-GB1	-----
YP_005861639.1	PDGTPKDETGYVPADQ-----VTDWSKIKSIIKTSSLSNN-----
WP_033688740.1	PTTVIPHVPGTTPKDFNGNPLKVPDPNDPSKGY-VPPTFENPTEDTQITYEKDTQKAKVT
WP_003097688.1	PTETLPYVPGFTPEDKDGN-----
WP_033689127.1	PTEVLPHYVPGFTPEDKDGNPLKVPDPTDPSKGYVVPNIPTDPSQDTVINYVANKANLVVK
WP_020903468.1	PTETLPYVPGFTPEDKDGNPLKVPDPNDPTKGYEVPSIPTNPGEDTPINYVANKANLVVK
WP_000287308.1	PTETLPYVPGFTPEDKDGNPLKVPDPNDPTKGYEVPSIPTNPGEDTLINYVANKANLVVK
WP_033685380.1	DKPVLPHYVPGMTPKDKDGNPLKVPDPNDPTKGYEVPSVPTNPGEDTPINYVKDKQKAKTT
WP_033681547.1	PTEVVPYVPGYTPKDKDGNPLKVPDPNDPTKGYEVPNVPTNPGEDTPINYVANKANLVVK
WP_023948409.1	PTDVLPHYV-----
WP_021874218.1	PDEPVPTIPGYTPSTP-----TVTPTDPGKDTVPYNPVVPKAD-----QKAVVN
KFL97016.1	PDEPVDPDIPGMTPTSTP-----TVTPEDPGKDTVPYNPVVPKAD-----QVAQVI
KFL95230.1	PDEPVDPDIPGMTPTSTP-----TVTPEDPGKDTVPYNPVVPKAD-----QVAQVI

Figure A2. Continued.


```

WT-GA
WT-GB1
YP_005861639.1 -----DRSDRLIFTGIDPN-----LVNDAGKTGYI-----STG
WP_033688740.1 YV---VEGTGTVLHTDNLEGKSGEPIEYS---TVTKLAELKALGYDLVTDGFTTATDKN
WP_003097688.1 -----
WP_033689127.1 YV---DENGKDLIPSETTEGKVGD--EYT---TTG---KVI PGHLLVRVEG--ESK GK
WP_020903468.1 YV---DENGKELQPTETKEGKVG D--DYS---TSG---KVITGYVLDRVEG--EAK GK
WP_000287308.1 YV---DENGKELLPTETKEGKVG D--DYS---TSGKV---ITGYVLDRVEG--EAK GK
WP_033685380.1 FV---DEKGNPIPGVDAITEEGSDT PLTKEAEVKAKIKEL ENKGYELVSNTY--PEG GK
WP_033681547.1 YV---DEKGDLLPAETTEGKVG D--EYA---TSGKV---IKGYVLVRVDG--EAK GK
WP_023948409.1 -----
WP_021874218.1 YVDA-DEDNKLITSSGDLTGKAGK KIDYS----TSSTIEDLINKGYVLVNDGF--PKDAT
KFL97016.1 YRDVQDGANKQLATSGDLTGKSG SEISYS---TADQIKKLINQGYVLKNDGF--PAGAV
KFL95230.1 YRDVQDGANKQLATSGDLTGKSG SEISYS---TADQIKKLINQGYVLKNDGF--PAGAV

WT-GA
WT-GB1
YP_005861639.1 FYS DTT-----
WP_033688740.1 YDKD TKVDQS FVVTVKPHVEPIK PVDPENP-----NDPNKPNPGD PIDPNPDGPKWTE D
WP_003097688.1 -----
WP_033689127.1 IGKDGSTVTYVYKPI-----
WP_020903468.1 IGT DGT TVTYVYKPLG-SWIPNIPGQPTNPIKYPNDPTDPTKPGQPTETLPYVPGYTPKD
WP_000287308.1 IGT DGT TVTYVYKPL-----
WP_033685380.1 FDKDKD TDQEFKVT LKERVVPV-----TPDQPKTPGAPVDPNNPDGPKYPAG
WP_033681547.1 IGKDGSTVTYVYKPL-----
WP_023948409.1 -----
WP_021874218.1 YDND DNTTQTYTVVFKHGTVPV-----TPTNPGKPGEPINPNPDGPKWPDG
KFL97016.1 FDND DSKNQVFYVDFI HGQAPV-----NPDNPHEGIDPS-----
KFL95230.1 FDND DSKNQVFYVDFI HGQAPV-----NPDNPHEGIDPS-----

WT-GA
WT-GB1
YP_005861639.1 -----
WP_033688740.1 LIKKIDTTRHVNRTITYVNEKGEEVAKK-----VTDKVTFTREGKINTVTGE-ITYGDW
WP_003097688.1 -----
WP_033689127.1 -----
WP_020903468.1 GNGQ-PLKPVDPQDPTKGYVVPNIPTDP-----SQD TVINYEANKAKLVVKYVDENGKD
WP_000287308.1 -----
WP_033685380.1 LEEK-DLNKTVIRTITYVYEDGT PVLNEDGTPKVVTQEAKFTREAKVNLVTGE-VTYGDW
WP_033681547.1 -----
WP_023948409.1 -----
WP_021874218.1 TGEN-SIDKTVTRTITFVDSNGKEVSSP-----VEQSVHFTATGVIDKVTGKWTPLSW
KFL97016.1 -----QYEKTVTEKVHYVGAGDKTPADN-----VQNSKWTRTLTIDVTGKVVENGQY
KFL95230.1 -----QYEKTVTEKVHYVGAGDKTPADN-----VQNS-KWTRTLTIDVTGKVVENGQY

```

Figure A2. Continued.

```

WT-GA
WT-GB1
YP_005861639.1
WP_033688740.1 TAKDGD-----TTFDKVESPVVKGYL---KDAKQKEVAATTGLTADSKDENIKVVYVP
WP_003097688.1
WP_033689127.1
WP_020903468.1 LIPSETTEGKVGDEYTTSGKVI PGHLLVRVDGDAKGKIGTEGSTVTYVYKPI-----
WP_000287308.1
WP_033685380.1 TPAQDL-----AEVKSPVVKGYLA-----DKVTVP TTKVTADSKDTKEVVYTKP
WP_033681547.1
WP_023948409.1
WP_021874218.1 SPDQSI-----DGRNV PFVDGYHVVSIDKDGNGLTNVKGVTLTHDDSSYKVTVTYVQ
KFL97016.1 TTDWSIAKGEKTVYDQVSTPVIDGYHA-----DKREVPATAVTQDD--IEVTVTYKP
KFL95230.1 TTDWSIAKGEKTVYDQVSTPVIDGYHA-----DKREVPATAVTQDDI--EVTVTYKP

WT-GA
WT-GB1
YP_005861639.1 ~KPF-ISSLAV-YNSSTVANKVKPANITITGEANINFK-----LQYTDE
WP_033688740.1 VGKVTITVPPGVTPTPTVTDTPYENVPGE PGKVVPPSPTKPQDPQDPNSPKVPVIPHIPG
WP_003097688.1
WP_033689127.1 ~GSW-IPNIPG-QPTTPIKYPNDPQDPTKPGQ--PTEV-----LPYVPG
WP_020903468.1 ~GSW-IPNIPG-QPTNPIKYPNDPQDPTKPGQ--PTEV-----LPYVPG
WP_000287308.1 ~GSW-IPNIPG-QPTNPIKYPNDPTDPTKPGQ--PTET-----LPYVPG
WP_033685380.1 LGSW-IPNIPG-QPTNPIKYPNDPADPTKPGSDKPV-----LPYVPG
WP_033681547.1 ~GSW-VPNIPG-QPTNPIKYPNDPTDPTKPGQ--PTEV-----VPYVPG
WP_023948409.1
WP_021874218.1 NGKI-IPVDPNGKPIPNVPTPQYPTDPTDPSKVV PNEP-----VPTIPG
KFL97016.1 NGKI-IPTDPSSNPIPNVNPNTYPTDPTDPTKVV PDQP-----VPEVPG
KFL95230.1 NGKI-IPTDPSSNPIPNVNPNTYPTDPTDPTKVV PDEP-----VPNIPG

WT-GA
WT-GB1
YP_005861639.1 NGQ-----LQ TINLPDLSTSYNLAQNNTMLTEQEAI ELANKNAASSI
WP_033688740.1 TTPVVPKDP TKPISP DNPLVPLTPVDNDPTKGYEVP VP TDPSTDT P-ITYVTDKQKAI
WP_003097688.1 -----PLKPVDPDTPTKGYVVPNIPTDPSQDTV-INYVANKAKLV
WP_033689127.1 FTP-----EDKDG NPLKPVDPKDP SKGYVVPNIPTDPSQDTV-INYVANKAKLV
WP_020903468.1 FTP-----EDKDG NPLKPVDPDTDPSKGYVVPNIPTDPSQDTV-INYVANKAKLV
WP_000287308.1 FTP-----EDKDG NPLKPVDPNDPTKGYEVPSIPTNPGEDTP-INYVANKANLV
WP_033685380.1 HTP-----VDGNGQPLKPVDPKDP SKGYEVPNVPTNPGEDTP-INYVANKANLV
WP_033681547.1 YTP-----KDKDG NPLKPVDPNDPTKGYEVPSVPTNPGEDTP-INYVANKANLV
WP_023948409.1
WP_021874218.1 YTP-----S-----TPTVTPTDPGKDTVPVYNPVEAKQGSVQVIFHDDTTNTT
KFL97016.1 MTP-----STPTVTPEDPGKDTVPVYNPVKNPDKVTTV-----EGKQI
KFL95230.1 MTP-----S-----TPTVTPEDPGKDTVPVYNPVVPAKDQAAVVNYVDADNNN

```

Figure A2. Continued.

```

WT-GA
WT-GB1
YP_005861639.1 PANYEIKSATLQSGGKT-----WQTDAPGPTPV
WP_033688740.1 TNFVDEKGVVSTPVVDEGDSGANFTKSKVDEVTKTIEKLEKAGYRVVKNFSPKDTDRV
WP_003097688.1 VKYVDENGKDLIPAETTEGKVGDEYTTSGK-----VIPGYVLRVDGEAKGKIGT
WP_033689127.1 VKYVDENGKDLIPSETTEGKVGDEYTTTGK-----VIPGYLLVRVEGEAKGKIGK
WP_020903468.1 VKYVDENGKDLIPAETTEGKVGDEYTTTGK-----VIPGYLLVRVDGDAKGKIGK
WP_000287308.1 VKYVDENGKELLPTETKEGKVDDYSTSGK-----VITGYVLRVEGEAKGKIGT
WP_033685380.1 VKYVDENGKELLPSETTEGKVGDEYATSGK-----VITGYVLRVEGEAKGKIGE
WP_033681547.1 VKYVDEKGDLLPAETTEGKVGDEYATSGK-----VIKGYVLRVDGEAKGKIGK
WP_023948409.1 -----
WP_021874218.1 IPDVGYNSGSQKAGTKVDYTTTKSISDL-----EVKGYVYVSTDGTIPTEITA
KFL97016.1 VHFVDGDNGNT-----PLRDPNTQTHEFKITNGVPD
KFL95230.1 TIITSSGNLTGKAGSRIDYSTKTTIADL-----ENKGYVLVNDGFPADATFDN

WT-GA
WT-GB1
YP_005861639.1 FGGQV---QYFYNNATVLLQAVPIQRTLTYQVIDENDPS-NPVTIQPNT-----
WP_033688740.1 FDKDKSVDQIFTVTVAERIIIPVTPGKPVDPNDPNLPKNPDGTPVTPSTPEPGKPVFPGBP
WP_003097688.1 DGSTV---TYVYTPLGSWVPNIPGQPTSP--IKYPNDPT-DPTKPGSDK-----
WP_033689127.1 DGSTV---TYVYKPIGSWIPNIPGQPTNP--IKYPNDPA-DPTKPGSDK-----
WP_020903468.1 EGSIV---TYVYKPIGSWIPNIPGQPTNP--IKYPNDPQ-DPTKPGSDK-----
WP_000287308.1 DGTTV---TYVYKPLGSWIPNIPGQPTTP--IKYPNDPQ-DPTKPGQPT-----
WP_033685380.1 NGTTV---TYVYKPLGSWVPNIPGQPTDP--IKYPNDPT-DPTKPGNDK-----
WP_033681547.1 DGSTV---TYVYKPLGSWVPNIPGQPTDP--IKYPNDPT-DPTKPGKDK-----
WP_023948409.1 -----
WP_021874218.1 DKNIT--VTVHMKHGTTTVTVPDKPKPGEPINPNPDGPK-WPDTTGKDNLSKTGTQTIH
KFL97016.1 ESSHT-----FTLVDPVIPGYVAEVKSAGGKTVTPDTPLAEV-----
KFL95230.1 DDSTTQVFTVVLKHGTVPVTPENPGKPGEPINSNDPDGPK-WPEGTDENSVKRTGTQTIH

WT-GA
WT-GB1
YP_005861639.1 -----
WP_033688740.1 NSPVWENTVKDLVTEKSATRTIKYVDRNGKEVSETRTETIKFTREAKVNIVTGEITYGEW
WP_003097688.1 -----
WP_033689127.1 -----
WP_020903468.1 -----
WP_000287308.1 -----
WP_033685380.1 -----
WP_033681547.1 -----
WP_023948409.1 -----
WP_021874218.1 YTGAGNNTPKDNVQSFTFTRTAVVDNVTGKVISTGAWNVTSHTFGNVNTPVVDGYHADKR
KFL97016.1 -----
KFL95230.1 YVGAGDKTPSDDVQTFDFTRKMVVDKVTGKVVDGGSWNVTSHTFGYKNTPVIDGYHADKR

```

Figure A2. Continued.

WT-GA	-----
WT-GB1	-----
YP_005861639.1	-----DLLNNG-----
WP_033688740.1	TTDRNDDIFNGYQVPVVKGYIAKAGDLESSTKDVQVTPDTIKDINETVIYDKLGSWIPNI
WP_003097688.1	-----
WP_033689127.1	-----
WP_020903468.1	-----
WP_000287308.1	-----
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----
WP_021874218.1	TAGNTTITPEDLNKTVTVNYTANG-----
KFL97016.1	-----TVVYHKVG-----
KFL95230.1	NAGGSVVT PDDLNKTVTVTYKQNG-----
WT-GA	-----
WT-GB1	-----
YP_005861639.1	-----QVITGNQGSNVPNDATTAYDAVKNLEAK-----
WP_033688740.1	PGTPTNPITYPNDPKDPTKPGTDKPKVPYVPGFI PVDPEGQPLKPVDPNDDPTK-----
WP_003097688.1	-----PVLPHYVPGYTPVDGNGQPLKPVDPNDDPTQ-----
WP_033689127.1	-----PVLPHYVPGHTPVDGNGQPLKPVDPNDDPTK-----
WP_020903468.1	-----PVLPHYVPGHTPVDGNGQPLKPVDPNDDPTK-----
WP_000287308.1	-----EVLPHYVPGFTPEDKDGNNPLKPVDPKDPSK-----
WP_033685380.1	-----PVLPHYVPGYTPKDKDGNNPLKPVDPNDDPTK-----
WP_033681547.1	-----PVLPHYVPGYTPKDKDGNNPLKPVDPNDDPTK-----
WP_023948409.1	-----
WP_021874218.1	-----KIIIPVDPNGKPIPNVPTPTYPTDPNDPTKVVPNEPV
KFL97016.1	-----KIVPVDPNGNPIPNVPTPSYTNDDPTDKVVPNEPV
KFL95230.1	-----KIVPVDPSGNPIPNVNPNTYPTDPTDPTKVVPDQPV
WT-GA	-----
WT-GB1	-----
YP_005861639.1	----GY--VISS--KSSTVPTIFGP-----
WP_033688740.1	----GY--EVPDVPGDPTQDTPINYIP-----
WP_003097688.1	----GY--EVPNVPNDDPTKDTPIINYVP-----
WP_033689127.1	----GY--EVPDIPTNPGEDTPINYIP-----
WP_020903468.1	----GY--ISPDIPTNPGEDTPINYIP-----
WP_000287308.1	----GY--VVPNIPTNPGEDTPINYIP-----
WP_033685380.1	----GY--EVPNVPTNPSEDTPINYIP-----
WP_033681547.1	----GY--EVPSVPTNPGEDTPINYVP-----
WP_023948409.1	-----
WP_021874218.1	PTIPGYKPSVPTVTPSDPGKDTVPYAP-----
KFL97016.1	PAITGKTPDKTSVTPVDPTKDTPVVYKN-----
KFL95230.1	PEVPGMTPSTPTVTPEDPGKDTVPYNPVKNPDKVTTVEGKQIVHFVDGDNGNTPLRDPN

Figure A2. Continued.

WT-GA	-----
WT-GB1	-----
YP_005861639.1	-----DNT-----
WP_033688740.1	-----
WP_003097688.1	-----
WP_033689127.1	-----N-----
WP_020903468.1	-----NVTP----NG
WP_000287308.1	-----NVTP----NG
WP_033685380.1	-----
WP_033681547.1	-----
WP_023948409.1	-----
WP_021874218.1	-----QTPVTPN--
KFL97016.1	-----
KFL95230.1	TQTHEFKITNGVPDESSHTFTLVDVPVPGYVAEVKSAGGKTVTPDTPPLAEVTVVYHKVG
WT-GA	-----
WT-GB1	-----
YP_005861639.1	-----ALTIYVTH-----
WP_033688740.1	----KDPTPNPTYP-----GTPFAPTP-----
WP_003097688.1	----APQPNPTP-----APTPKPEP-----
WP_033689127.1	----SPKPNPTYP-----GTPFAPTP-----
WP_020903468.1	DQNGYTPQPKPQPEQVVTTYVDENGKDIAPSE-----
WP_000287308.1	DQDGYTPQPKPQPEQVVTTYVDENGKDIAPSE-----
WP_033685380.1	----NSPKPNPTYP-----GTPFAPTP-----
WP_033681547.1	-----
WP_023948409.1	-----
WP_021874218.1	--IPVTPNEPSTPTT-----PDTSAPTP-----
KFL97016.1	NEVPATPNSQKAVVNFIDVNTGKLIKTSGLS-----
KFL95230.1	KIVPVDPNGNPIPNVPTPSYTNDPTDPTKVVPNEPVPAITGKTPDKTSVTPVDPTKDTPV
WT-GA	-----
WT-GB1	-----
YP_005861639.1	-----KTINV-----
WP_033688740.1	-----KPEPKPEPKPEPKPEPE-----
WP_003097688.1	-----KPEPKPE-----
WP_033689127.1	-----KPEPKP-----
WP_020903468.1	-----KGAQAPKGISGYEYVTTTKDPN
WP_000287308.1	-----KGAQAPKGISGYEYVTTTKDPN
WP_033685380.1	-----KPEPKP-----
WP_033681547.1	-----
WP_023948409.1	-----
WP_021874218.1	-----HGEDVPVTPNEP-----
KFL97016.1	-----GRPGEDINKLYSSAEVIKQLEEAG
KFL95230.1	VYKNNEVPATPNSQKAVVNFIDVNTGKLIKTSGLSGRPGEDINKLYSSAEVIKQLEEAG

Figure A2. Continued.

WT-GA	-----
WT-GB1	-----
YP_005861639.1	-----STPDQWPSTSTDADKV-----SLSKITRITITVEGLP
WP_033688740.1	-----TPQPVTADDGDNNNGN-----NNGTP-STPAQPAAP
WP_003097688.1	-----TPQPVTADNGDNNNGN-----NNETP-TTPAQPAAP
WP_033689127.1	-----EPAPVPSTPETPEQPVAPV-----QPEQP-TTPTQPAVP
WP_020903468.1	GNLVHHYKKVATPQPVPSTPETPEQPVAPV-----QPEQQ-TNPNQPAVP
WP_000287308.1	GNLVHHYKKVATPQPVPSTPETPEQPVAPV-----QPEQP-TTPTQPAVP
WP_033685380.1	-----EPAPVPSTPETPEQPVAPM-----QPEQP-TNPNQPAVP
WP_033681547.1	-----NPREV-EKPAKPAQP
WP_023948409.1	-----
WP_021874218.1	-----DTPAPAPHGEKPEEPDRPA-----PAPHAP-KAPTAKGNN
KFL97016.1	YEVVYNAFDGDGVTKYFDDDDNTTQQFTVALKLKEKAKTPYPVVPAPETPAKEPEAPA EK
KFL95230.1	YEVVYNAFDGDGVTKYFDDDDNTTQQFTVALKLKEKAKTPYPVVPAPETPAKEPEAPA EK
 WT-GA	 -----
WT-GB1	-----
YP_005861639.1	T-----AVEGTTQTVTFTRTAVVDEVTKGVIGYVDPSTDSQTITDGDNAWTSVNNT-WSA
WP_033688740.1	S-----TPQYMDGQ-----REL PNTGTEDHASLAALGLL-GAL
WP_003097688.1	S-----TPQYMDGQ-----REL PNTGTEDNASLAALGLL-GVL
WP_033689127.1	T-----PAETSVATDSATQTATPKYVDGQ-----KELPNTGTEANASLAALGLL-GAL
WP_020903468.1	A-----PAETSVATDSATQPATPKYVDGQ-----KELPNTGTEANASLAALGLL-GAL
WP_000287308.1	T-----PAETSVPTDSATKPATPKYVDGQ-----KELPNTGTEANASLAALGLL-GAL
WP_033685380.1	A-----PAETSVATDSATQPATPKYVDGQ-----KELPNTGTEANASLAALGLL-GAL
WP_033681547.1	S-----KQETPKYVEGQ-----KELPNTGTEANASLASLGLL-GAL
WP_023948409.1	-----
WP_021874218.1	T-----PVKENKTVPTAA-PVVKNEQTPE-----AELPQTGEKND SAAAILGATAGMI
KFL97016.1	VSRPEQPVKQNVSVPTPQKPVEKKTNNKK-----EVL PQTGADNNEAASILGAVATAI
KFL95230.1	VSRPEQPVKQNVSVPTPQKPVEKKTNNKK-----EVL PQTGADNNEAASILGAVATAI
 WT-GA	 -----
WT-GB1	-----
YP_005861639.1	FTP-----
WP_033688740.1	SGFGLIARKKKREDEE-
WP_003097688.1	SGFGLVARKKKED---
WP_033689127.1	GGFGLLTRKKKED---
WP_020903468.1	GGFGLLARKKKED---
WP_000287308.1	GGFGLLARKKKED---
WP_033685380.1	GGFGLLSRKKKED---
WP_033681547.1	GGIGLLTRKKKED---
WP_023948409.1	-----
WP_021874218.1	GLIGLLGVKKKHSEN-
KFL97016.1	GMTSLIGAKRRKKDDK
KFL95230.1	GMTSLIGAKRRKKDDK

Figure A2. Continued.

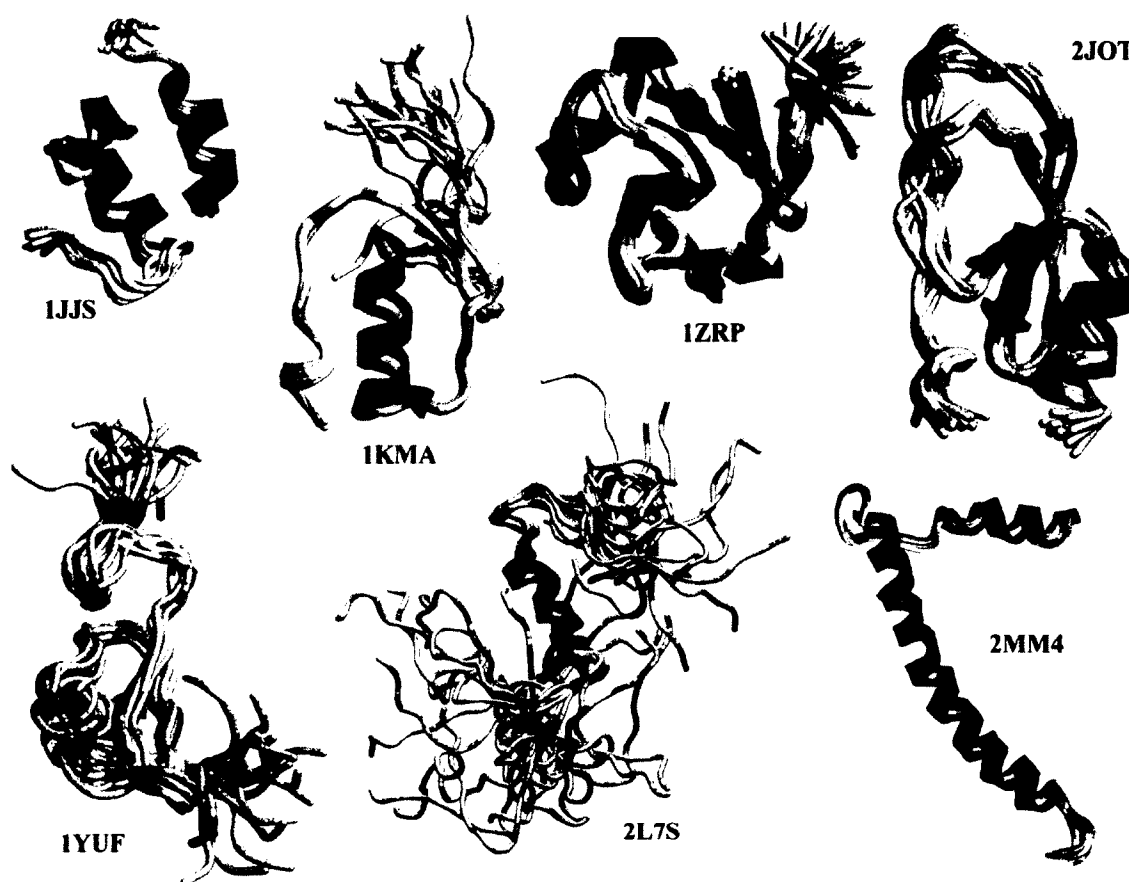


Figure A3. Structure of 7 small globular proteins. PDB codes are presented with each structure. The α -helices are colored in pink and β -strands colored in yellow. The loop regions as well as the N- and C-terminal are colored grey. All images were created with RasMol (Ver. 2.7.2.1.1).

APPENDIX II

METHODS FOR CLONING, EXPRESSION AND PURIFICATION OF WT-GB1 AND VARIANTS

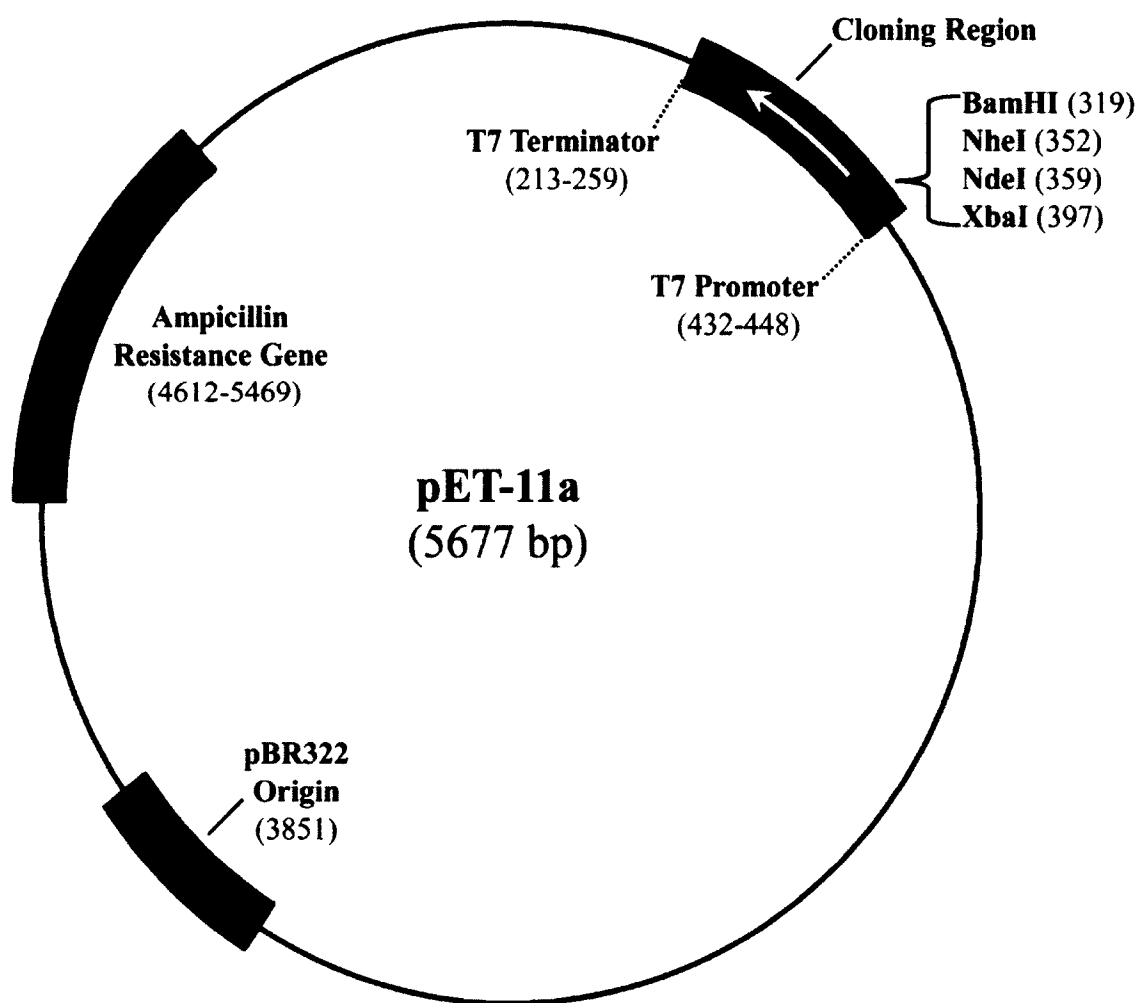


Figure A4. Diagram of the pET-11a vector map. Schematic shows the ampicillin and cloning region with the T7 promoter and terminator. Image was redrawn from <http://www.genomics.agilent.com/en/home.jsp>.

APPENDIX III

NMR METHOD TO STUDY LONG-RANGE INTERACTION IN GB1

Table A1. GB1 residue side-chain solvent accessibility

Residue			Residue		
		% Relative Accessibility			% Relative Accessibility
1	Met	65.2	29	Val	66.4
2	Thr	56.2	30	Phe	3.2
3	Tyr	4.5	31	Lys	48.8
4	Lys	36.1	32	Gln	87.8
5	Leu	0	33	Tyr	41.8
6	Ile	48.7	34	Ala	2.6
7	Leu	2.4	35	Asn	94.9
8	Asn	54	36	Asp	83.9
9	Gly	0	37	Asn	50.4
10	Lys	84.7	38	Gly	100
11	Thr	98.9	39	Val	3.2
12	Leu	33.3	40	Asp	100
13	Lys	63.6	41	Gly	22.4
14	Gly	95.8	42	Glu	99.1
15	Glu	55.5	43	Trp	20.6
16	Thr	50.9	44	Thr	80.9
17	Thr	52.9	45	Tyr	25.5
18	Thr	31.9	46	Asp	66.8
19	Glu	79.7	47	Asp	74.9
20	Ala	14.4	48	Ala	90
21	Val	94.2	49	Thr	70
22	Asp	59.9	50	Lys	36.5
23	Ala	37.7	51	Thr	17.9
24	Ala	82.3	52	Phe	3.4
25	Thr	41.5	53	Thr	36.8
26	Ala	0.5	54	Val	0.1
27	Glu	47.2	55	Thr	56.2
28	Lys	72.1	56	Glu	13.4

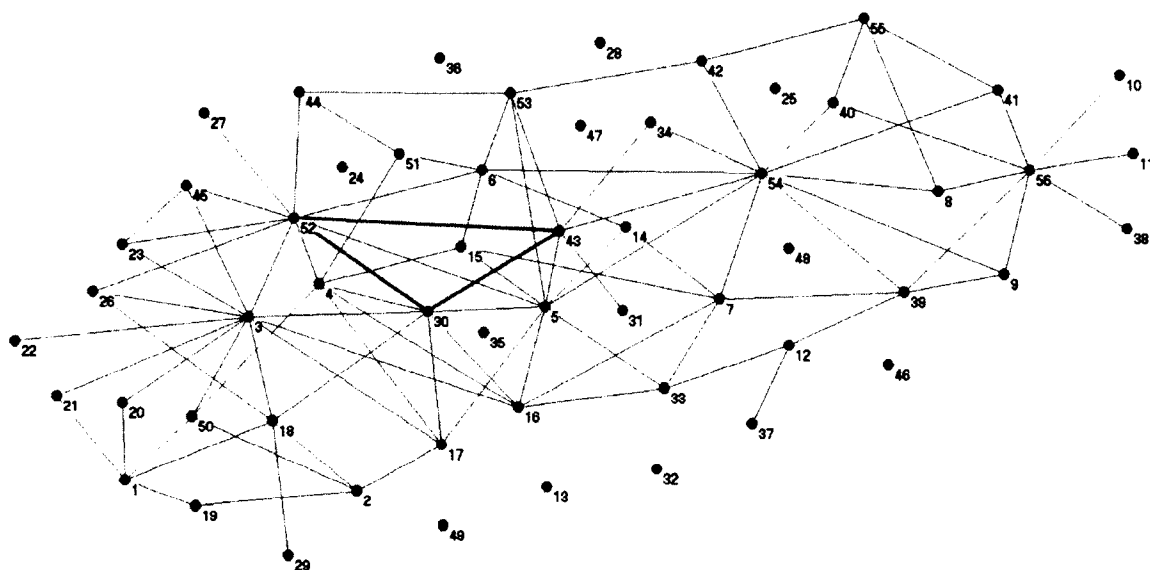


Figure A5. Network schematic of long-range interactions in GB1. Red circles indicate residues and blue lines represent contact between residues that are within 5 Å and greater than 7 residues apart from each other in the sequence. Orange lines indicate interactions of interest between Phe30, Phe 52 and Trp 43.

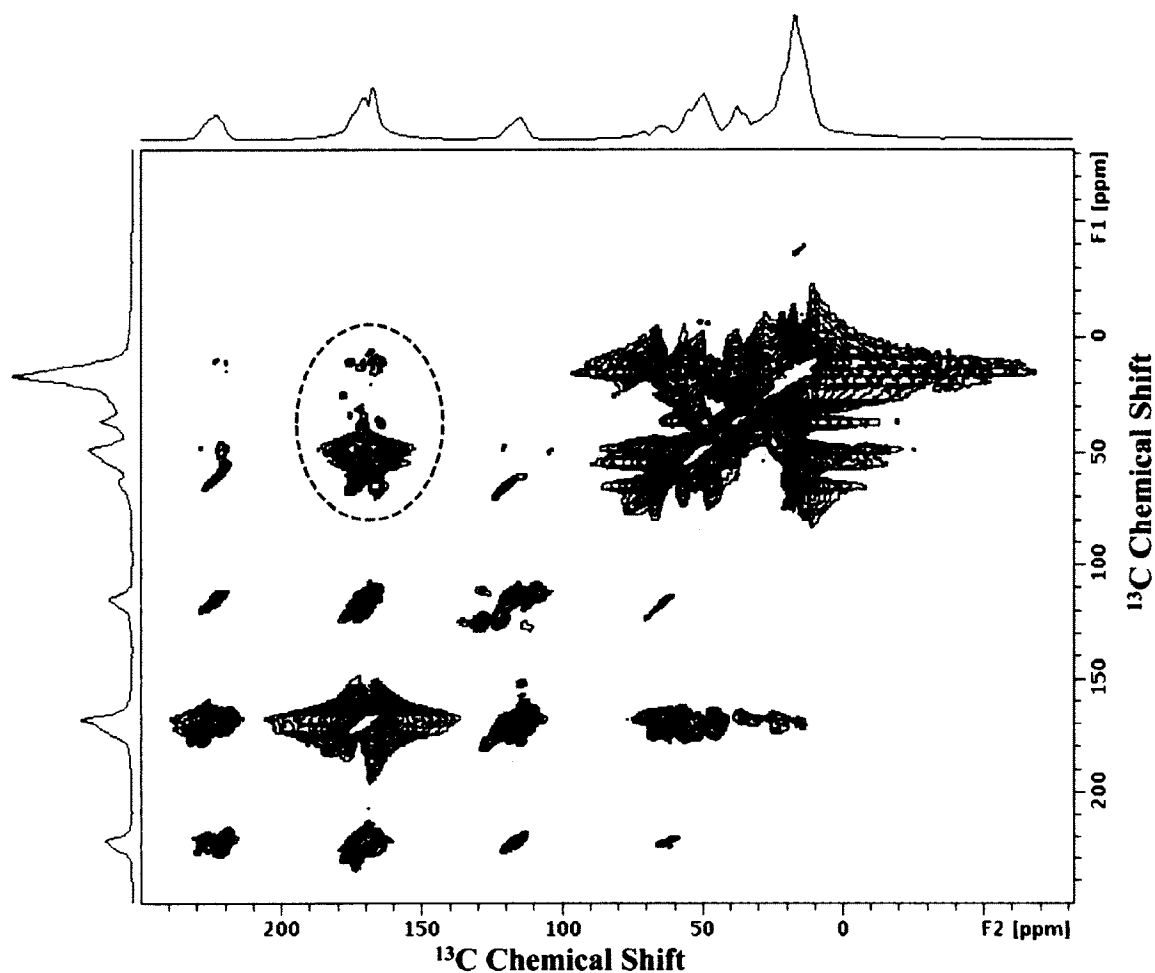


Figure A6. 2D ^{13}C - ^{13}C -DARR experiment on uniformly labeled WT-GB1. A dry sample of ~36 mg uniformly labeled WT-GB1 was packed into a 4 mm rotor and analyzed on a Bruker Avance 750 MHz MAS-NMR spectrometer located at the College of William and Mary. The sample was spun at 10 kHz. Outlined in a dashed red circle are representative CRPK.

Table A2. All long-range interactions within 5 Å found in GB1

Residues			Residues			Residues		
A	B	# Contacts	A	B	# Contacts	A	B	# Contacts
1	18	3	5	43	6	20	26	3
1	19	15	5	52	19	23	45	6
1	20	16	5	53	3	23	52	6
1	21	1	5	54	5	26	52	9
1	50	5	6	14	9	27	52	16
2	17	3	6	15	7	30	43	9
2	18	11	6	51	4	30	52	11
2	19	16	6	52	9	31	43	30
2	50	4	6	53	22	34	43	3
3	16	3	6	54	9	34	54	1
3	17	11	7	13	13	38	56	4
3	18	19	7	14	23	39	54	3
3	20	12	7	15	5	39	56	20
3	21	3	7	16	1	40	54	1
3	22	17	7	33	6	40	55	2
3	23	12	7	39	2	40	56	19
3	26	8	7	54	11	41	54	4
3	30	9	8	54	13	41	55	10
3	45	8	8	55	25	41	56	7
3	50	25	8	56	9	42	53	3
3	52	6	9	39	2	42	54	10
4	15	19	9	54	1	42	55	17
4	16	13	9	56	15	43	52	6
4	17	11	10	56	24	43	53	16
4	30	11	11	56	1	43	54	32
4	50	7	12	33	3	44	51	3
4	51	24	12	37	6	44	52	11
4	52	14	12	39	3	44	53	32
5	14	3	16	30	8	45	51	18
5	15	9	16	33	9	45	52	36
5	16	24	17	30	8			
5	17	2	18	26	3			
5	30	31	18	29	1			
5	33	2	18	30	15			

Degree long-range contacts between residues A and B with a contact distance cutoff of 5 Å and residue sequence spacing of 6 residues. Comprehensive list of 1025 long-range interactions found in the structure of GB1. List represents potential interactions to be monitored by folding-freezing stopped-flow coupled with NMR.

APPENDIX IV

REPRINT PERMISSIONS

Permission to Reprint The Following Figures was Obtained from Publishers:

Figure 4	Figure 26
Figure 5	Figure 27
Figure 7	Figure 28
Figure 9	Figure 29
Figure 10	Figure 33
Figure 11	Figure 58
Figure 14	Figure 59
Figure 15*	Figure 75
Figure 16	Figure 76
Figure 17	Figure 77
Figure 18	Figure 78
Figure 19	Figure 79
Figure 21	Figure 80
Figure 24	Figure 81
Figure 25	

*Copyright (2005) National Academy of Sciences, U.S.A.

Electronic copies of the permissions are on file and maintained on the ODU Department of Chemistry and Biochemistry K-Drive as a permanent record.

VITA

Jason Charles Collins

Department of Chemistry and Biochemistry
Alfriend Chemistry Building
Old Dominion University
Norfolk, VA 23529

Education

2015 (GPA: 3.86/4.0): Ph.D. Biomedical Sciences - Old Dominion University, Norfolk, VA

2008 (GPA: 3.12/4.0): B.S. Biochemistry - Old Dominion University, Norfolk, VA

Publications

Collins J.C., Greene L.H., Biophysical Analysis of the Transition of an all α -Helical Greek-Key Protein into Amyloid Fibrils Composed of β -Sheet Structure. *Protein & Peptide Letters* (2012) 19:982-990.

Collins J.C., Greene L.H., Conversion of α -helical proteins into an alternative β -amyloid fibril conformation. *Bionanoimaging: Protein Misfolding and Aggregation* (2012) 485-501.

Munyanyi A., Collins J.C., Greene L.H., Inhibition of β -synuclein fibril formation *in vitro* by α -synuclein. (2014). (Submitted)

Collins J.C., Bedford J., Greene L.H., Conservation and Network Analysis of the ($4\beta+\alpha$) Fold of the Immunoglobulin-Binding B1 Domain of Protein G to Elucidate the Key Determinants of Structure, Folding and Stability. (2015). (Submitted)

Munyanyi A., Bedford J., Collins J.C., Sori N. and Greene L.H., Origins of the Synucleins: Orphans or Superfamily Members? (2015). (Submitted)

Conference Presentations (Oral)

Collins J.C., Greene L.H., Exploratory Approaches To Visualize Non-Covalent Long-Range Interactions In Proteins Using Solid-State Nuclear Magnetic Resonance. SERMACS, Nashville, TN, October 16-20, 2014.

Collins J.C., Greene L.H., Conversion of an All α -Helical Greek-key Protein into Highly Ordered Amyloid Fibrils Composed of β -sheet Structure. VAS, Norfolk, VA, May 24–27, 2012.

Awards

CIBA Fellowship (\$5,000), Old Dominion University, May 2013 – August 2013.

University Fellowship (\$15,000), Old Dominion University, Aug 2011 – May 2012.

FASEB (MARC) Program Travel Award (\$1,850), The Protein Society, 2012.

FASEB (MARC) Program Travel Award (\$1,650), The Protein Society, 2011.